# Cancel Culture and Information Transmission*

Davide Cianciaruso†, Ilan Guttman‡, and Iván Marinovic§

**Abstract**

We study information transmission in a setting where informed senders sequentially communicate to a decision maker. The innovation of our model is in considering an opponent who always prefers to maintain the status quo (due to disagreement with the decision maker). The opponent may impose costs on a sender, by cancelling the sender, in order to deter future senders from influencing the decision-making process away from the status quo. The equilibrium obtains one of four mutually exclusive patterns. Surprisingly, sometimes the larger the disagreement between the senders and the opponent, the more information transmission is elicited in equilibrium.

**Keywords:** Information transmission, cancel culture, experimentation, reputation.

**JEL Classification:** D72, D82, D83, G20.

*We thank Cyrus Aghamolla, Jeremy Bertomeu (discussant), Seung Lee, Pierre Liang (discussant), Hans Frimor, Russell Lundholm, Kevin Smith, as well as seminar participants at the DAR&DART Accounting Theory Seminar, UCLA, NYU, Stanford Accounting Summer Camp, Universidad de Chile, the University of Southern Denmark, the 13th Accounting Research Workshop, and the 11th University of Alberta Accounting Research Conference for helpful feedback.

†New Economic School.

‡NYU Stern.

§Stanford GSB.

# 1 Introduction

Learning is a collective and cumulative process. The evolutionary success of the *Homo Sapiens* as a species is given in part by its ability to accumulate knowledge across generations. At each point in time, human beings can use the stock of knowledge created by prior generations and develop it further.

Due to the sequential nature of learning, any impediments to knowledge creation today hinder knowledge production tomorrow. Yet, the flow of ideas is far from being frictionless: new ideas often face strong opposition, and more so the more monolithic the orthodoxy is.

The instinct to censor new ideas and punish dissent has been prevalent throughout the history. It is told that Pythagoras, the Greek mathematician, punished his disciple Hippasus to death by drowning for asserting the existence of irrational numbers, contrary to the beliefs of the Pythagorean school. The Catholic church sentenced Galileo to house arrest for life for challenging Aristotle's geocentrism.

A recent tendency to suppress debate is often referred to as "cancel culture." In a New York Times column, Ross Douthat defines cancellation as "an attack on someone's employment and reputation by a determined collective of critics, based on an opinion or an action that is alleged to be disgraceful and disqualifying." Leaving aside the arguments around each particular incident, such a practice is believed to cause reluctance to state certain opinions. For instance, a letter published in the Harper's Magazine in 2020, which was signed by more than 150 intellectuals from across the political spectrum, recognizes that "the result [of censoriousness] has been to steadily narrow the boundaries of what can be said without the threat of reprisal. We are already paying the price in greater risk aversion among writers, artists, and journalists who fear for their livelihoods if they depart from the consensus [...]."

In this paper, we offer a theoretical model to study the implications of the cancel culture phenomenon on information transmission. Specifically, we model learning as a collective task where a sequence of speakers/senders may be punished/cancelled for voicing their beliefs. In this context, does greater disagreement necessarily hinder communication? We find that the answer is no.

Our model features three types of players. *(i)* There is a decision maker who, based on all the available information at the end of the game, makes a decision (high/low action). The decision maker aims at taking the action that is closest to the state of nature. *(ii)* There are two senders, each of whom obtains a conditionally independent noisy private signal of the state of nature. The signal is binary, say "high" (H) or "low" (L). Senders report their signals sequentially. *(iii)* There is an opponent, with a strong preference, or beliefs, for the low action (status quo). In each period, the opponent has the ability to punish/cancel a sender if the sender reports a high signal. We often refer to this preference misalignment as "bias," but it may capture difference of opinions.[1] Our model is agnostic about which agents have the "right" opinion. Our focus is on information transmission and not on welfare, which would have required one to take a stand on which of the agents' opinions (if any) is correct.

We focus on the interesting case where, in order to sway the decision maker away from the status quo (i.e., the action that is consistent with prior beliefs) there is a need for sufficient support for such a decision by the senders. In our setting with two senders, both of them must credibly disclose a high signal to induce a deviation of the decision maker from the status quo. In other words, given the decision maker's priors, a single (high) signal is not enough to deviate from the status quo. This assumption creates an interdependence between the senders' strategies that captures the collective nature of learning.

In the model, speech, namely, publicly voicing private information, is useful to the decision maker but can be costly to the speaker. One such cost comes from the risk of being targeted by opponents who have incentives to block communication of certain information, and are willing to impose costs (cancelling) on dissenting speakers (e.g., career, reputation, etc.). A speaker can report the signal truthfully or lie about it incurring no direct cost (cheap-talk). In particular, a speaker may report L when, in fact, the private signal is H in order to avoid the risk of being

[1]The potential disagreement between the decision maker and opponent can be due to various reasons. For example, the source of disagreement may be the opponent's status quo bias, i.e., a preference for the low action that is consistent with the prior belief; disagreement about the interpretation of the speakers' private information and/or the prior beliefs in the sense that the expectation of state of nature given an information set is different by the decision maker than by the opponent; disagreement about the precision of the private signals of the senders, or the precision of the prior beliefs.

canceled. Lying is thus a form of self-censoring.

On the one hand, being cancelled is costly to the speaker, potentially exerting a deterrent effect on future speakers and affecting the society's ability to learn about the state of nature. On the other hand, the act of cancelling the speaker may also be costly to the opponent. While some opponent types may gain a direct benefit from punishing the speaker, others incur direct costs from doing so (henceforth, we refer to the former type of opponent as a "zealot"). The opponent privately observes his type – his cost of cancelling a speaker.

We assume that speakers have a natural preference for truthfully communicating their signals in order to maximize the learning by the decision maker. However, speakers are uncertain whether the opponent will attack them if they challenge the status quo, since the speakers do not perfectly know the opponent's type, captured by the opponent's cost/benefit from attacking a speaker.

Absent any disagreement/difference of opinion among agents, or other source of some incentive misalignment, there would not be any friction in the learning process. We study the more interesting (and often realistic) setting which is characterized by sufficient incentive misalignment between the opponent and the decision maker and senders. In our main analysis, we assume that the agents disagree about the prior probability that the state is high (and the agents agree to disagree). In particular, we assume that the prior of the opponent is such that even given two high realizations of the noisy signals the opponent still believes that the state is more likely to be low than high. As such, even given two high realizations of the signals, the opponent still prefers the low action to be taken by the decision maker over the high action. In contrast, both the decision maker and the senders' prior beliefs are such that given one positive realization and one negative realization of the signal they prefer the low action, however, given two positive signal realizations they prefer the high action to be taken. We allow for disagreement also between the senders and the decision maker, as long as the above holds.[2] All of our model's results are qualitatively the same if we assume that there is a bias in the objective function of the opponent relative to that of the decision maker and the senders – a form of preference misalignment similar to the standard

[2]An alternative source of disagreement that yields similar results is disagreement among agents regarding the precision of the senders' private signals, rather than the precision of prior beliefs.

bias between the sender and the receiver in cheap-talk models (Crawford and Sobel, 1982). While difference of opinion slightly complicates the analysis relative to a bias in the objective function, we believe that it is a better fit to the cancel culture application, as agents that decide to cancel speakers may truly believe that their action is taken to better with society as a whole.

We begin by noting that cancellations can only emerge in equilibrium if there is a positive probability that the opponent is a zealot. Otherwise, the opponent would never cancel a speaker, no matter how strong his preference is for the low action. The reason stems from backward induction: if all opponent types have a positive cost from cancelling, an opponent does not have an incentive to cancel the last speaker, because at that point all the information has already been conveyed. But then, canceling the first speaker is not a credible threat to deter the second speaker from being truthful. In sum, without a positive probability for a zealot opponent, cancellation and self-censorship cannot arise in equilibrium (given that the game has deterministic finite periods.)

Next, we discuss the results under the assumption that there is a positive probability that the opponent is a zealot. Our first main result is an equilibrium taxonomy that depends on the senders' cost of being cancelled and on the opponent's extent of disagreement among the agents – or equivalently, the extent of the opponent's preference towards the status quo (aka, opponent bias or disagreement) . In particular, depending on the cost of being cancelled, the equilibrium can take on one of four patterns, which we refer to as: *(i) No Deterrence (ND)* equilibrium, in which the first sender always reports truthfully and, conditional on a high report by the first sender, also the second sender reports truthfully; *(ii) Partial Deterrence (PD)* equilibrium, in which the first sender always reports truthfully and, even if the first sender was cancelled, the second sender still truthfully reports a high signal with some probability; *(iii) Full Deterrence (FD)* equilibrium, in which the first sender always reports truthfully and the second sender truthfully reports a high signal only if the first sender was not cancelled; and *(iv) No Information Transmission (NIT)* equilibrium, in which both sender types always report a low signal (pooling equilibrium).

Our taxonomy provides a clear characterization of the mutually exclusive parameter values for which each of the four equilibrium patterns prevails. When deciding whether to truthfully report a

high signal, a sender weighs the benefit from the probability that he will induce a high action (when both signals are high) against the expected cost of being cancelled. The cases of ND and NIT are very intuitive. When the cost of being cancelled is sufficiently low (high) a sender's benefit from truthfully reporting a high signal is higher (lower) than the expected cost from being cancelled. The intuition for FD and PD cases is less trivial. In the following, we elaborate first on FD (which is the simpler of the two) and then on PD.

*Full Deterrence (FD) Equilibrium.* The FD pattern occurs for moderate senders' cost of being cancelled. This equilibrium entails deterrence, whereby the first speaker reports his signal truthfully, but if the first speaker is canceled by the opponent, the second speaker self-censors and never reports a high signal. A cancellation in the first period induces the second speaker to update upwards the probability that the opponent is a zealot and the likelihood that he himself will be cancelled upon reporting H. This effect induces opponents, even those who have a positive cost of canceling, to cancel the first speaker, in order to build a reputation for "toughness," and thus influence the transmission of information and the decision making process.[3]

*Partial Deterrence (PD) Equilibrium.* The PD also occurs for relatively moderate cost of being cancelled that are smaller than in the FD case. In this equilibrium, following cancellation of the first speaker, a second speaker with a high signal randomizes between truthfully reporting and reporting a low signal. Note that if the second speaker were to tell the truth with probability one following a first period cancellation, then cancellations would not play a deterrence role, and hence only zealots would be willing to cancel the first-period speaker. In turn, this means that observing a first period cancellation would reveal perfectly that the opponent is a zealot, and that the cancellation of second period dissenters is certain. But this, in turn, would hinder the incentive of the second speaker to dissent. In other words, mixing by the second sender ensures that cancellation of the first sender has a deterrence effect, and thus some opponents who are non-zealots also have an incentive to cancel first-period speakers. In turn, this lowers the probability that the second speaker will be cancelled conditional on first period cancellation (as the opponent may not be a zealot) and allows

[3]For some parameter values (mainly when the bias of the opponent is extremely high), FD does not exist, and an increase in the cost of being canceled will move the equilibrium directly from PD to NIT.

the second speaker to be truthful with positive probability.

Under PD, an increase in the cost of being cancelled results in a lower probability that the second sender truthfully reports a high signal, and hence the amount of information that is conveyed in equilibrium decreases. Interestingly, depending on the parameter values, small increases in the cost of being cancelled can lead to a spiral of self-censorship or a discrete drop in the amount of information transmission (from PD to NIT). Under PD, following a first period cancellation, a second sender observing a high signal is indifferent between truthfully reporting it or not. When the cost of being cancelled increases, the only way to preserve the indifference of the second sender and maintain a mixed strategy is if the probability that the second sender will be cancelled decreases, in such a way that the expected payoff from truthful reporting remains constant. Because in the second period only zealots cancel the sender, for the posterior probability of cancellation perceived by the second sender to go down, a higher fraction of opponents who are not zealots must cancel the first sender. This in turn can happen only when the probability that the second sender truthfully reports a high signal (conditional on a first period cancellation) decreases, thereby strengthening the opponent's incentive to cancel the first sender. As the cost of being canceled keeps increasing, it eventually reaches a point where the cost of the first sender from truthfully reporting a high signal exceeds his expected benefit from truthful reporting (note that the likelihood of implementing a high action also decreases due to the decreased likelihood that the second sender truthfully reports a high signal). Once the first sender no longer truthfully reports the high signal, also the second sender ceases to truthfully report it, because a single dissenting signal, by itself, does not sufficiently move the beliefs of the decision maker to change his status quo action. In other words, for parameter values that yield the Partial Deterrence (PD) equilibrium, as the cost of being cancelled increases, eventually the amount of information transmission drastically reduces to zero, leading to NIT.

Surprisingly, under PD an increase in opponent disagreement (bias) leads to an increase in information transmission. An increase in opponent's disagreement (bias) boosts the opponent's incentives to cancel the first speaker, other things equal. But the opponent's propensity to cancel

the first speaker cannot change, or else the posterior beliefs of the second sender that he will be cancelled would change, thereby violating his indifference between truth-telling and self censoring. Hence, after an increase in opponent's disagreement (bias), the second sender must increase the probability of being truthful conditional on first period cancellation, to offset the effect of an increase in disagreement (bias) on the opponent's incentive. These two effects must exactly offset each other, such that the opponent's behavior remains unchanged. In summary, since the first sender always reports truthfully and the second sender increases his propensity to truthfully report high signals, the overall information transmission increases in opponent's disagreement (bias).

Since under the FD, the information transmission decreases in the magnitude of the opponent's disagreement (bias) (and obviously is unaffected in NIT and ND), the effect of the opponent's disagreement (bias) on information transmission is non-monotone whenever both a PD and a FD equilibrium exist.

**Related literature.** To the best of our knowledge, the cancel culture phenomenon has not been studied in economics, with the exception of Lowery and Carvalho (2021). They consider a Bayesian Persuasion setting where the opponent may be "woke," which means he assigns zero probability to some states of nature. The woke opponent cancels/punishes the sender if the sender's experiment proves that the true state is one to which the opponent assigns zero prior probability. The risk of being cancelled leads the sender to avoid experiments that may prove the opponent wrong. The threat of triggering undesirable actions by other parties, and the consequences of such a threat on sender's ex ante choices, are also present in Ben-Porath, Dekel, and Lipman (2018). They consider an ex-ante project choice and subsequent voluntary disclosure by a sender who seeks to maximize an observer's expectation of the project outcome. They show that the presence of a challenger, who instead seeks to minimize the observer's expectation, can induce the sender to choose a low-risk project when he would otherwise choose an excessively risky one. Antic, Chakraborty, and Harbaugh (2020) study a related problem in which two senders engage in communication under the scrutiny of an observer with different incentives who can object to the senders' decision. The two

senders can use a communication protocol that allows them to implement their preferred choice in spite of the scrutiny of the observer.

Acemoglu and Wolitzky (2023) explore a related model of conflict between two parties who face uncertainty about each other's preferences and actions. They characterize the conditions under which peaceful equilibria can arise and show that, in certain scenarios, conflict can escalate into spirals or cycles. However, their analysis does not center on conflict as a barrier to information transmission, which is the primary focus of our study.

Our paper builds upon several strands of the economics literature. First, it relates to the reputation literature. In the chain store paradox game (see Selten, 1978; Kreps and Wilson, 1982), a long-term player builds reputation for being tough, aiming to influence the behavior of a sequence of other players. Similarly, in our model, the opponent might decide to cancel today's speaker to dissuade future speakers, who are afraid of being canceled as well, from sending the same message. Morris (2001) considers a model where senders are concerned about their reputation and some messages are ex ante more likely to be released by "bad" types. Reputation concerns reduce the sender's propensity to be truthful to avoid being associated with the bad type. This is interpreted as the sender trying to be politically correct. Loury (1994) examined the concept of self-censorship as a means to evade reputational damage and its broader social impacts, including the implications of such actions for learning and information transmission.

Second, our paper is related to the literature on learning, specifically, on how the actions of economic agents affect this process. Learning in markets goes back to Keynes and Hayek (see Hayek, 1945). In "The use of knowledge in the society," Hayek argues that the key advantage of markets relative to socialism stems from the ability of prices to aggregate disperse information, and guide the resource allocation process. In herding models (see Bikhchandani, Hirshleifer, and Welch, 1998, 1992, 2016; Ottaviani and Sørensen, 2006) economic agents, who act sequentially and are affected by others' actions, may ignore their private information, thereby leading to rational cascades and strong inefficiencies.

Our paper is also related to the experimentation literature (see Keller and Rady, 1999; Keller,

Rady, and Cripps, 2005; Aghion, Bolton, Harris, and Jullien, 1991). Ours is a model of *collective* experimentation. By sending 'forbidden' messages and observing the opponent's reaction, senders elicit information about the type of the opponent, thus helping subsequent senders to better understand the costs and benefits of controversial speech. Chen, Du, Stocken, and Wang (2024) analyze a two-period game featuring a principal and two agents, all long-lived. In each period, agents can engage in misconduct to obtain a private benefit, at the expense of the principal. The principal can discipline the agents, undoing the misconduct, but at a privately known cost. In this dynamic model, the principal has an incentive to enforce good behavior in the first period, in order to build a reputation for strictness and deter future misbehavior. At the same time, the agents have an incentive to misbehave, in order to elicit information about the principal's cost of enforcement. When the enforcement actions are public, there are information externalities from the agent's peer.

While our main focus is on the implications of cancel culture on information transmission, we believe that our model may also provide pure theoretical contributions to the extant literature. We build on a sequential cheap talk model, but we introduce another economic agent besides the senders and the receiver/decision maker. Such an agent – the opponent – has an incentive to block communication of some realizations in order to affect the ultimate decision. The novelty of the introduction of such an agent is two-fold. First, unlike most cheap talk models, strategic communication does not emerge because of the preference misalignment between senders and receivers. Instead, it appears because a third party may punish senders for stating what he disagrees with. Second, one can consider our model as a collective experimentation game with a *strategic* bandit. Senders are initially uncertain how likely they will be attacked by the opponent, which depends on the opponent's private attack cost and his endogeonus strategic choice. The first sender's disclosure policy can generate information for the second sender. The presence of the opponent leads to new tensions and predictions, such as the non-monotonicity of informativeness as a function of disagreement (bias).

# 2 Model

We study a strategic communication model that involves three types of players: i) Two informed senders who can publicly reveal their private information about the state of nature in a sequential fashion, ii) a strategic opponent who can impose costs (cancelling) on senders, and iii) a decision maker who makes a single decision $a \in \{0, 1\}$ at the end of the game.

We study a setting with two periods, $t = 1, 2$, where in each period, a privately informed sender (aka speaker) issues a report. After observing the report in a given period, an opponent (O) may publicly punish the speaker. We refer to this punishment as "canceling" the speaker. At the end of the game, based on the information that the speakers have revealed, the decision maker (DM) chooses between two possible actions. Figure 1 depicts the sequence of events. The action of the decision maker affects the payoffs of all the players, as described below. The opponent considers canceling the current-period's speaker in order to influence future speakers, and thereby affect the DM's action.

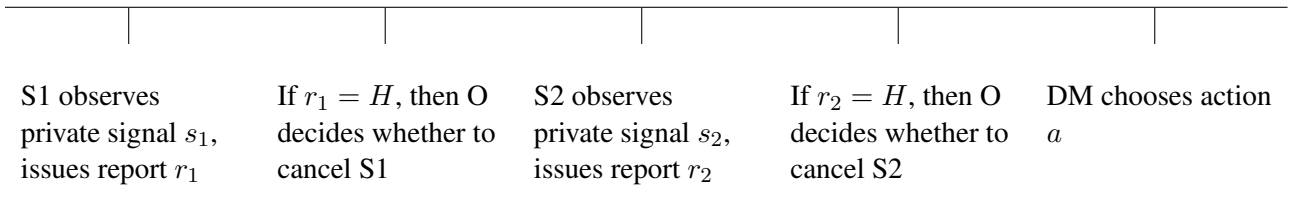



Figure 1: Timeline.

We next provide a detailed description of the model's setting and assumptions. In Section 2.1, we further discuss and motivate our assumptions. Agents' payoffs depend on a binary state of nature, $\tilde{\theta} \in \{\theta_L, \theta_H\}$, where $\theta_L < \theta_H$.[4] To capture difference of opinions in the simplest way, we assume that agents disagree on the prior probability that the state is high (e.g., Che and Kartik, 2009). Specifically, agent $i \in \{S, O, DM\}$ believes that $\Pr_i(\tilde{\theta} = \theta_H) = q_0^i$, where the subscript "S" here stands for either of the senders. Note that the model does not require taking a stand on

[4]Throughout, we use a tilde to denote random variables (e.g., $\tilde{\theta}$) and we drop it to denote their realizations (e.g., $\theta$).

which player (if any) has the correct beliefs.

We assume that agents prefer actions that match the state (according to their beliefs). To capture this preference, we assume that each agent incurs a loss given by the quadratic distance between the state and the action. If agent $i$ expects DM to take action $a$, and $i$'s belief that the state of nature is high is $q^i$, then

$$V(a; q^i) \equiv \mathrm{E}_{q^i}\left(-(a-\tilde{\theta})^2\right) = -q^i(a-1)^2 - (1-q^i)(a)^2$$

denotes the expected loss for agent $i$. When agent $i$'s belief is $q^i$, his differential expected loss when he expects DM to take action $a = 1$ vs. $a = 0$ is

$$\Delta(q^i) \equiv V(1; q^i) - V(0; q^i) = 2q^i - 1. \tag{1}$$

The expression in (1) captures the idea that differences in beliefs translate into differences in preferences over the DM's action. With these preferences, agents prefer that the DM takes the action that is closer to their beliefs about the state of nature. In particular, an agent $i$ prefers the high (low) action if and only if $q^i > 1/2$ ($q^i < 1/2$).

**The Decision Maker (DM).** At the end of game, after observing all publicly available information (i.e., the reports of the two senders and the two canceling decisions of the opponent), the Bayesian DM chooses action $a \in \{0, 1\}$. DM chooses the action that minimizes the expected loss according to his beliefs.

**The Opponent (O).** We assume that, in the absence of information, O would want DM to choose the low action. That is, O's prior is $q_0^O < 1/2$, so that O has an innate preference for the low action. To create tension in the model, we assume that O's and DM's beliefs are such that following two positive signals O still believes that the state is more likely to be low – unlike the DM and the senders. This leads to an ex post objective misalignment following two high signals, and, therefore,

O has an incentive to block communication of high signals by speakers. Consistent with this, in each period, we allow O to cancel the speaker if the speaker reported $r_t = H$. We take O's prior $q_0^O$ relative to the priors of DM and speakers as a measure of disagreement: fixing DM's and speakers' priors, the lower $q_0^O$ is, the greater the disagreement among agents.

Depending on the opponent's type, O incurs a direct cost or a direct benefit from canceling a speaker. The cost/benefit of the opponent from canceling the sender, which is O's private information, is given by the realization of the random variable $\tilde{c}$. We assume that $\tilde{c}$ is continuously distributed according to the cumulative distribution function $F(c)$, with full support on the real line.[5] We allow the cost to be negative with positive probability to capture the fact that some opponents may have a direct benefit from canceling. Otherwise, in games with deterministic finite periods there would never be cancellation on the equilibrium path (see the discussion in Section 2.1). A negative $c$ (direct benefit from canceling speakers who report $H$) may represent a situation where some opponents are self-righteous and derive a moral satisfaction from punishing a particular type of speech. Henceforth, we will refer to such opponent types, with $c < 0$, as "zealots." The ex ante probability that O is a zealot, $F(0)$, is referred to as the "fraction of zealots." Note that a zealot always cancels a sender who reports $r_t = H$. In order to deter the second sender from reporting $r_2 = H$, in equilibrium, even opponents who are not zealots (i.e., even if $c > 0$) may cancel a sender who reports $r_1 = H$ in the first period. Opponents with sufficiently high cost of cancellation never cancel the speaker. As such, cancellation of the first sender increases the second sender's beliefs that O is a zealot, and hence increases the second sender's beliefs that, if he reports $r_2 = H$, he will be canceled too. We denote the decision of the opponent to cancel/punish (not cancel) the sender in period $t$ by $d_t = 1$ ($d_t = 0$). Thus, O's payoff, net of potential canceling costs, is

$$U_O(a, q, d_1, d_2) = V(a; q) - \sum_{t=1}^{2} d_t \cdot c.$$

[5] Assuming this range for $\tilde{c}$ avoids the uninteresting case where $c$ is always negative and O cancels the first sender with certainty.

**The Senders (Speakers).** We denote the speaker of period $t$ by $S_t$. Each of the two speakers privately observes an informative, but noisy signal, $\tilde{s}_t \in \{L, H\}$ of the state of nature. The speakers' private signals are independent conditional on the state of nature, and their distribution is characterized as follows:

$$\Pr_i\left(\tilde{s}_t = H|\tilde{\theta} = \theta_H\right) = \Pr_i\left(\tilde{s}_t = L|\tilde{\theta} = \theta_L\right) = \alpha \in \left(\frac{1}{2}, 1\right).$$

The parameter $\alpha$ is the probability with which the signal correctly identifies the state of nature, and represents the precision of $S_t$'s information, in a mean-preserving spread sense: the greater $\alpha$ is, the higher (lower) is the posterior conditional on high (low) signal.

In each period $t$, after having observed his private signal, speaker $S_t$ issues a public report, which is denoted by $r_t \in \{L, H\}$. The speaker is not confined to report his signal truthfully. For simplicity and parsimony, we assume that speakers bear no direct cost from misreporting their signal.

Being canceled is personally costly to the speaker: if canceled, the speaker bears a commonly known cost of being canceled, which is denoted by $k > 0$. Note that the speaker can always avoid such a cost by reporting $r_t = L$, whether truthfully or untruthfully. When deciding whether to report $r_t = H$ following a signal $\tilde{s}_t = H$, a sender considers the trade-off between the expected cost from being canceled and the expected benefit from being pivotal in inducing the decision maker to take the high action. Thus, the payoff of sender $S_t$ is

$$U_{S_t}(a, q, d_t) = V(a; q) - d_t k.$$

**Parameter values.** As a shortcut notation, let us denote the posterior beliefs of agent $i$ conditional on only the first signal and on both signals by

$$q^i_{s_1} \equiv \Pr_i\left(\tilde{\theta} = \theta_H|\tilde{s}_1 = s_1\right)$$
$$q^i_{s_1,s_2} \equiv \Pr_i\left(\tilde{\theta} = \theta_H|\tilde{s}_1 = s_1, \tilde{s}_2 = s_2\right),$$

respectively. Next, to create potential incentive misalignment/tension, we focus on the region of parameters that satisfy the following restrictions.

**Assumption 1.** DM's and speakers' prior beliefs about the state of nature are such that the posteriors

$$q_H^{DM}, q_H^S < \frac{1}{2} < q_{H,H}^{DM}, q_{H,H}^S.$$

**Assumption 2.** O's prior belief about the state of nature is such that the posterior

$$q_{H,H}^O < \frac{1}{2}.$$

Note that, for any agent $i$ and conditional on any information set, the posterior beliefs are a monotonically increasing function of the prior $q_0^i$. Therefore, Assumptions 1 and 2 can be equivalently stated in terms of bounds on the priors.[6]

Assumption 1 states that speakers' and DM's preferences are sufficiently aligned, and that they prefer the high action if and only if two high signals are received. The feature that only two consecutive high signals can sway DM towards the high action is descriptive of learning as a collective task, which requires multiple speakers to exert influence. By contrast, if only one high signal were enough for DM to choose the high action, then any non-zealot O would have no incentive to cancel the first speaker, S1. The reason is that, when the belief is in favor of the high action, O would actually want communication to take place, because an informative low report by S2 can make DM choose the low action instead.[7] Observe that Assumption 1 implies $q_0^{DM} < 1/2$, that is, without any information transmission DM would select the low action.

[6] Assumption 1 holds if and only if the initial beliefs satisfy

$$q_0^{DM}, q_0^S \in \left( \frac{(1-\alpha)^2}{1-2\alpha(1-\alpha)}, 1-\alpha \right)$$

and Assumption 2 if and only if

$$q_0^O \in \left( 0, \frac{(1-\alpha)^2}{1-2\alpha(1-\alpha)} \right).$$

These intervals are non-empty because $\alpha \in (1/2, 1)$ and, hence, for any $\alpha$ there exist priors that satisfy Assumptions 1 and 2.

[7] It is also possible that the canceling penalties for speakers are so small that full communication occurs even after first-period cancellation. Even in that case, O has no incentive to incur a positive cost given that it will not affect the DM's action anyway.

For what concerns the speakers, Assumption 1 ensures that they have sufficient incentives to truthfully report the high signal for some parameter values. For that to happen, a necessary condition is that speakers prefer the high action after two consecutive high signals. At the same time, speakers preferences should not be too tilted towards the high action relative to DM, otherwise communication would break down: an S2 who observed a low signal would lie, if reporting high led DM to choose the high action.

Assumption 2 states that O prefers the low action regardless of the signal realizations. Stated differently, the speakers' information is not precise enough, given O's prior, to make him prefer the high action even after two consecutive high signals. At the same time, note that O's payoff is state-dependent (just like the other players'), which implies that the strength of O's preference for the low action, and hence his willingness to cancel, is a function of his prior. This is an important feature of our model, as it allows us to perform comparative statics with respect to O's prior.

Assumptions 1 and 2 jointly require that there be sufficient disagreement between O, on one side, and DM and the speakers, on the other side. Without that, a strategic O would have no incentive to punish S1 for having reported high, in order to block further communication: O himself would want DM to take the high action after two consecutive high signals.

**Strategies and equilibrium.** Without loss of generality, we assume that in equilibrium players use the natural language (Ottaviani and Sorensen, 2006). That is, the report $r_t = H$ ($r_t = L$) is associated with a greater (smaller) posterior probability that the state of nature is high. We solve for perfect Bayesian equilibria of the dynamic game. Note that we focus on equilibria where DM plays pure strategies. As common in standard cheap-talk models, while a pooling equilibrium always exists (in which all senders issue a low report), when other equilibria exist, we focus on the most informative equilibrium.

**Informativeness.** We do not make welfare claims, as they depend on the agent's perspective. Instead, we focus on analyzing the transmission of information. To capture the amount of information transmitted in equilibrium, we calculate, for each realization of the speakers' signals, the prob-

ability with which they report truthfully. Formally, let $\tau(s_1, s_2) \equiv \Pr((\tilde{r}_1, \tilde{r}_2) = (s_1, s_2) | (\tilde{s}_1, \tilde{s}_2) = (s_1, s_2))$. It will be clear from the analysis that $\tau(L, L)$ and $\tau(L, H)$ are constant in all equilibria: $\tau(L, L) = 1$, because the low signal is always reported truthfully, and $\tau(L, H) = 0$, because S2 always reports low after S1 reports low. Also, $\tau(H, L)$ is a weakly monotone transformation of $\tau(H, H)$: if S1 does not truthfully report the high signal, then $\tau(H, L) = \tau(H, H) = 0$; and if $\tau(H, H) > 0$, then $\tau(H, L) = 1$, because S2 always truthfully reports the low signal. For these reasons, it suffices to focus our attention on $\tau(H, H)$ only.

In Appendix C, we demonstrate that the results are qualitatively similar if we alternatively defined informativeness as the ex ante expectation, calculated according to DM's prior, of a convex function of his posterior conditional on all public information at the end of the game. Thus defined, informativeness would capture the combined amount of information that the two speakers' report convey in equilibrium: informativeness would be greater if the speakers' reports increased the DM's posterior in a mean-preserving spread sense. Such a definition encompasses the special case where the goal is to minimize the mean squared error of the DM's posterior.

## 2.1 Discussion of assumptions

**Unraveling and zealots.** We assume that some opponents are zealots, in the sense that they have a negative canceling cost $\tilde{c}$ or, in other words, they gain direct benefit from punishing speakers who release a report $H$. In any game of our setting with finite and deterministic number of periods, if O's canceling cost were always positive, there would always be disclosure unravelling in equilibrium, that is, the speakers would always report truthfully. To see this, observe that opponents with positive canceling costs do not have an incentive to punish the last speaker, because at that stage DM's final belief is already formed. As a consequence, the last speaker is truthful. But because the behavior of the last speaker is unaffected by whether the penultimate speaker is canceled, O does not cancel the penultimate speaker either, who in turn is truthful, and so on. In sum, for cancellation to affect information transmission and induce censorship it is necessary to assume that with positive probability O enjoys canceling certain types of messages. This can be justified in

practice: zealots could derive a moral gain (e.g., virtue signaling) from punishing the speaker. Put differently, we do not need to assume that zealots are some form of sadistic individual, but rather a moral or religious individual committed to political correctness (see Morris, 2001).

**Disagreement.** Our model assumes that players have different prior beliefs. Alternatively, we could have imposed a common prior and assumed difference of opinions about the precision of speakers' information. Namely, Assumptions 1 and 2 are equivalent to assuming different beliefs about the precision of the sender's signal, such that $\alpha^{DM}, \alpha^{S} \in (\hat{\alpha}_1, \hat{\alpha}_2)$ and $\alpha^{O} < \hat{\alpha}_1$, for some cutoffs $\hat{\alpha}_1 < \hat{\alpha}_2$. That is, O believes that the precision of speakers' signal is lower than what DM and speakers think. A lower $\alpha^{O}$ would capture greater disagreement. Yet another equivalent formulation consists in assuming disagreement about the location of the state of nature on the real line. According to such a formulation, agent $i$ would believe that $\theta_L = b_i$ and $\theta_H = 1 + b^i$, where $b^{DM}, b^{S} \in (\hat{b}_1, \hat{b}_2)$ and $b^{O} < b_1$, for some cutoffs $b_1 < b_2$. That is, O's preferences lean more towards the low action than DM's and speakers'. These two alternative versions yield qualitatively the same results because the agent's actions are driven by their preferences and, eventually, differences in beliefs boil down to differences in preferences (Morris, 1995). Details are available upon request.

**Multiple opponents.** Our model assumes a single opponent, such as an influential individual, an association or an organization. This single opponent trades off his impact on DM's eventual action with the cost/benefit of cancelling speakers. While a setting with a single, or a leading, opponent is descriptive of certain situations, sometimes cancellation ensues from the uncoordinated action of a multitude of small agents. If opponents' payoff were only a function of the state and DM's action, free-riding could occur with a continuum of opponents, as none of them would be pivotal. However, there are reasonable instances where even small opponents would have incentives to engage in costly cancellation. For instance, if they received a benefit $B > 0$ from taking active part in a successful cancellation campaign (and no benefit if either the campaign fails or if they abstained from it). We may assume that the probability of successful cancellation is linearly

increasing in the fraction of opponents who participate in the campaign, so that each individual opponent is more keen to take part if he expects more other opponents to do the same. Additionally, in the spirit of global games, we may assume that each opponent $i$ receives a private noisy signal about the common cost/benefit of cancelling speakers, $\tilde{c}$. The signal is $\tilde{x}_i = \tilde{c} + \sigma\tilde{\varepsilon}_i$, where $\tilde{\varepsilon}_i$ is a noise component and $\sigma > 0$. Using standard solution techniques (e.g., Morris and Shin, 2003), in the limiting equilibrium as the noise vanishes (i.e., $\sigma \to 0$), opponents follow a cutoff strategy such that they all participate, and cancellation is successful with certainty, if $c \leq \hat{c} = \frac{1}{2}B$, and otherwise no one participates, and cancellation fails. Speakers' updating following cancellation (lack thereof) are as in our main model (analyzed below). That is, upon cancellation (lack thereof), speakers know that the opponents' cost of cancelling is below (above) a cutoff. Last, if we assume that the benefit $B$ is endogenous and proportional to the difference in opponents' expected utility from cancellation vs. no cancellation, then the multiple-opponent model would be mathematically equivalent to the main model (up to a transformation of the distribution of $\tilde{c}$). The details of this extension are available upon request.

# 3 Analysis

## 3.1 Preliminary observations about the equilibrium

Before deriving the equilibria of the game for the various parameter values, we make three initial observations regarding the players' equilibrium behavior. The first observation relies on the parametric assumptions above regarding the posteriors $q^i_H$ and $q^i_{H,H}$, whereas the second and the third observations are more general. First, there is no equilibrium in which S2 reports truthfully after S1 reported $L$. This occurs because, if S2 reports $H$, he will trigger canceling whenever O is a zealot. At the same time, S2 will still not be able to sway the DM away from $a = 0$ (note that if S2 believes that S1 truthfully reported $L$, he would not even want the action to be $a = 1$). Thus, S2 self-censors to avoid cancellation if S1 reported $L$.

Second, observe that in the second period only zealots will cancel S2, because at that point

cancellation cannot affect the information transmission process any longer. Hence, if S2 knew that O has a positive canceling cost (i.e., is not a zealot), then S2 might have preferred to report truthfully .

Third, in any equilibrium, O plays a threshold strategy in the first period. That is, there exists a threshold level of the opponent's direct cancellation cost/benefit, which we denote by $\hat{c}$, such that O cancels S1 following a report $H$ if and only if O's canceling cost is $c < \hat{c}$.

Given the cheap-talk nature of this game, there is always a pooling equilibrium, where no information transmission takes place. In such an equilibrium, the DM ignores the speakers' messages and both speakers report $L$ regardless of their signals. Henceforth, we will refer to the pooling equilibrium as the no-information transmission equilibrium (NIT).

Next, we study the more interesting possibility of equilibria with information transmission.

## 3.2 No Deterrence Equilibrium

In this section, we consider the possibility of an equilibrium in which the first speaker, S1, truthfully reports and the second speaker truthfully reports following a high report by S1 – even if S1 was cancelled. In this equilibrium, DM always has sufficient information to choose his optimal action, given the joint information set of both speakers. We refer to this equilibrium as the "No Deterrence" (ND) equilibrium.

Following $r_1 = H$ and $d_1 = 1$, S2 truthfully reports $H$ only if

$$V(1; q^S_{H,H}) - \Pr_S\left(\tilde{d}_2 = 1 | \tilde{d}_1 = 1\right) k \geq V(0; q^S_{H,H}).^8 \tag{2}$$

Equation (2) is S2's incentive compatibility constraint. The left-hand side is S2's expected payoff from truthfully reporting $H$. By reporting truthfully, S2 ensures that DM takes the optimal action given S2's posterior, $q^S_{H,H}$, but S2 also incurs a positive expected cost from being cancelled -

[8]To ease the notational burden, we simply write $\Pr_S\left(\tilde{d}_2 = 1 | \tilde{d}_1 = 1\right)$ instead of $\Pr_S\left(\tilde{d}_2 = 1 | \tilde{r}_1 = H, \tilde{d}_1 = 1, \tilde{r}_2 = H\right)$, on the grounds that canceling can only occur if the speaker reports $H$. Similarly, below we write $\Pr_S(\tilde{d}_1 = 1)$ instead of $\Pr_S(\tilde{d}_1 = 1 | \tilde{r}_1 = H)$ for the probability that O cancels S1 in the first period.

$\Pr_S\left(\tilde{d}_2 = 1|\tilde{d}_1 = 1\right) k$. The right-hand side is S2's payoff from falsely reporting $L$. If S2 reports $L$ instead, he avoids the expected cost of being cancelled but gets disutility from the action of the DM – which will be $a = 0$. In the ND equilibrium, only zealots are cancelling speakers. Therefore, O's equilibrium cutoff for canceling in the first period is $\hat{c}_{ND} = 0$. Hence, first-period cancellation in the ND equilibrium implies that S2 expects certain cancellation if he reports $H$. That is, $\Pr_S(d_2 = 1|d_1 = 1) = 1$ and (2) boils down to $k \leq \Delta(q^S_{H,H})$, where $\Delta(\cdot)$ was defined in (1).

S1 anticipates that if he truthfully reports $H$, then S2 will report truthfully for any realized signal. Therefore, in the ND equilibrium, S1 is willing to truthfully report $H$ only if the following S1's incentive compatibility constraint holds:

$$\begin{aligned} &\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)\, V(1; q^S_{H,H}) + \Pr_S(\tilde{s}_2 = L|\tilde{s}_1 = H)\, V(0; q^S_{H,L}) - \Pr_S\left(\tilde{d}_1 = 1\right) k \\ &\geq V(0, q^S_H). \end{aligned} \tag{3}$$

The left-hand side of (3) is S1's expected payoff from truthfully reporting $H$. In that case, (under the conjectured ND) S2 will also report his signal truthfully and DM will take the optimal action conditional on both signals. The right-hand side is S1's payoff from misreporting. If S1 reports $L$, then S2 will always report $L$, because of the assumption that a single high signal is not enough to induce DM to take the high action. Therefore, in the event that S2's signal is $H$, DM will take the suboptimal action $a = 0$. If S1 reports $H$, his probability of being canceled is equal to $\Pr_S\left(\tilde{d}_1 = 1\right)$. In the ND equilibrium, in which O's cutoff in the first period is $\hat{c}_{ND} = 0$, this probability is $\Pr_S\left(\tilde{d}_1 = 1\right) = F(0)$. The incentive-compatibility condition of S1 (3) hence simplifies in the ND equilibrium to $k \leq \frac{\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)}{F(0)} \Delta(q^S_{H,H})$.

Overall, we conclude that the ND exists if

$$k \leq \min\left\{ \frac{\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)}{F(0)} \Delta(q^S_{H,H}), \Delta(q^S_{H,H}) \right\} \equiv k_1. \tag{4}$$

$k_1$ in (4) thus denotes the highest speakers' cost of being canceled for which the ND equilibrium

exists.

Generically, one of the two IC constraints (3) and (2) is redundant, depending on whether $\frac{\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}{F(0)} \gtrless 1$. On the one hand, S2 bears a higher expected cost of being canceled in the ND equilibrium than S1 (sure cancellation vs. $F(0)$). On the other hand, S2 derives a greater benefit from being truthful, because he is always pivotal for DM's decision, whereas S1's disclosure of $H$ will make a difference only in the event that S2 also discloses $H$. One can see that S1 has stronger incentives for truthtelling than S2 when either the ex ante fraction of zealots, $F(0)$, is relatively low or when S1's conditional probability that S2's signal is also $H$, $\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)$, is relatively high.

If condition (4) is violated, then the ND does not exist. In the following subsections, we solve for equilibria with partial information transmission. In general, there cannot be an equilibrium with partial information transmission where S1's report is uninformative, because S2 cannot be pivotal without the information that S1 provides. However, there can be equilibria where S2 misreports with positive probability. As long as there is a relatively high probability that S2 will be truthful if needed, S1 will have an incentive to be truthful. The rest of the equilibrium taxonomy depends on the fraction of zealots, $F(0)$, and the level of disagreement, as captured by the distance among priors. These are the two variables that jointly determine the intensity of O's opposition to the high action.

## 3.3 Full Deterrence Equilibrium

Let us identify the conditions under which an equilibrium exists where S1 is truthful, but S2 is truthful only if S1 reported $H$ and O did not cancel him. We label this equilibrium as the "Full Deterrence" (FD), because O's decision to cancel S1 has the ability to censor S2, thereby deterring further information transmission.

In this context, the opponent has control over DM's actions, because by canceling S1 he shuts down S2 and ensures $a = 0$. However, for some O types canceling is too costly, and so they may prefer to forego canceling S1 and run the risk that DM will implement $a = 1$.

In the FD, the equilibrium cutoff $\hat{c}_{FD}$ for O's cost of canceling makes him indifferent between canceling S1 for having reported $H$ and not canceling S1. Canceling S1 induces $a = 0$ for certain, whereas not canceling S1 leads to the risk that $a = 1$ – if S2 also reports $H$. This trade-off leads to the following indifference condition for O's cancellation decision of S1:

$$V(0; q_H^O) - \hat{c}_{FD} = \Pr_O\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right) V(1; q_{H,H}^O) + \Pr_O\left(\tilde{s}_2 = L | \tilde{s}_1 = H\right) V(0; q_{H,L}^O), \quad (5)$$

where on the left-hand side (resp., right-hand side) of (5) we have O's payoff from canceling (not canceling). By simplifyng the expression, we can solve explicitly the equilibrium cutoff,

$$\hat{c}_{FD} = (1 - q_H^O)(1 - \alpha) - q_H^O \alpha. \quad (6)$$

Note that the equilibrium cutoff in (6) is such that $\hat{c}_{FD} > 0$. A positive cutoff cost means that some O types who find canceling personally costly are willing to cancel S1 when doing so deters S2 from releasing signal $H$. Canceling today allows O to build a reputation for toughness, thereby deterring information transmission tomorrow. The following trade-off determines the value of the cutoff. The benefit of canceling S1 is avoiding a loss when the state is low and S2 incorrectly observes and reports a high signal. From O's perspective, this event occurs with probability $(1-q_H^O)(1-\alpha)$. The cost of canceling S2 is the loss when the state is high, S2 observes a high signal, but does not report it because he fears cancellation. This event occurs with probability $q_H^O \alpha$. Naturally, increases in O's prior decrease the resulting cutoff and, in turn, the ex ante probability that S1 is canceled. This finding is intuitive: when O's prior is higher, the posterior $q_H^O$ is also higher. Consequently, O is relatively less concerned about the DM taking the high action when the state is low and relatively more concerned about the DM taking the low action when the state is high.

Taking O's cutoff $\hat{c}_{FD}$ as given, S2 is unwilling to be truthful when observing a high signal after first-period cancellation if

$$V(1; q_{H,H}^S) - \Pr_S\left(\tilde{d}_2 = 1 | \tilde{d}_1 = 1\right) k \leq V(0; q_{H,H}^S), \quad (7)$$

where $\Pr_S\left(\tilde{d}_2 = 1|\tilde{d}_1 = 1\right) = F(0)/F(\hat{c}_{FD})$ is the conditional probability that S2 is canceled given that S1 was canceled. The inequality in (7) boils down to

$$k \geq \frac{F\left(\hat{c}_{FD}\right)}{F\left(0\right)}\Delta(q^S_{H,H}) \equiv k_2. \tag{8}$$

In essence, this condition states that, for S2 not to truthfully reveal $H$, the expected cost of being canceled has to be greater than the benefits of inducing the optimal action.

Going backwards to the first period, S1 has an incentive to truthfully report $H$ if

$$\begin{aligned}\Pr_S\left(\tilde{d}_1 = 1\right)\left[V(0, q^S_H) - k\right] + \Pr_S\left(\tilde{d}_1 = 0\right)&\begin{bmatrix}\Pr_S\left(\tilde{s}_2 = H|\tilde{s}_1 = H\right)V(1, q^S_{H,H}) \\ + \Pr_S\left(\tilde{s}_2 = L|\tilde{s}_1 = H\right)V(0, q^S_{H,L})\end{bmatrix} \\ \geq V(0, q^S_H).&\end{aligned} \tag{9}$$

The probability of S1 being canceled is $\Pr_S\left(\tilde{d}_1 = 1\right) = F\left(\hat{c}_{FD}\right)$. The left-hand side of (9) is S1's payoff from truthfully reporting $H$. In the FD, if S1 reports $H$ and is canceled ($\tilde{d}_1 = 1$), then future communication breaks down and DM will choose $a = 0$. If instead S1 reports $H$ but O does not cancel him, then S2 will be truthful and the DM will take his (and the speakers') optimal action under symmetric information. The right-hand side of (9) is S1's payoff from reporting $L$, which leads to the DM choosing $a = 0$ for certain. Simplifying S1's IC condition yields

$$k \leq \frac{\left(1 - F\left(\hat{c}_{FD}\right)\right)\Pr_S\left(\tilde{s}_2 = H|\tilde{s}_1 = H\right)}{F\left(\hat{c}_{FD}\right)}\Delta(q^S_{H,H}) \equiv k_3. \tag{10}$$

We can think of truthful disclosure by S1 as an experimentation attempt to elicit information about O's type. In particular, experimenting to learn whether or not O is a zealot. If S1 discloses $L$, no such information can be gathered, because O never cancels low reports. But when S1 discloses $H$, O has to make a decision whether to cancel S1. If O chooses not to cancel, then speakers infer that O is not a zealot, because his cost of canceling must be $c > \hat{c}_{FD} > 0$. Therefore, S2 can safely be truthful in the second period knowing that he will not be canceled. If instead O does cancel S1 who

disclosed $H$, then speakers still revise their beliefs about O's type but some residual uncertainty remains. As a consequence, S2 enters the second period knowing that he runs the risk of being canceled, if he also discloses $H$. S1's cost of experimenting is given by O's potential decision to cancel S1. On the other hand, S1's benefit from experimenting is that, after no cancellation in the first period, S2 will be truthful and DM will take S1's desired action when both speakers' signals are high. This trade-off gives rise to the cutoff value $k_3$ for S1's cost of being canceled.

Summing up, for the FD to exist we need

$$k \in [k_2, k_3], \tag{11}$$

where $k_2$ and $k_3$ are defined in (8) and (10), respectively. In other words, speakers' cost of being canceled, $k$, has to be large enough to deter truthful reporting by S2 after first-period cancellation, but low enough for S1 to have an incentive to be truthful, even though there is a chance that S2 might be deterred. There are certain parameter configurations for which the interval in (11) is empty and, hence, the FD cannot exist for any value of speakers' cost $k$. For the FD to exists, S1's IC constraint needs to be slacker than S2's, in the sense that, for the same cost of being canceled, $k$, S1 is willing to truthfully disclose $H$ when S2 is unwilling to do so following cancellation of S1.

As it turns out, we can pinpoint whether the interval is empty to how intense is O's opposition to the high action. In the model, the intensity of O's opposition is captured by two parameters. First, the ex ante probability that O is a zealot $F(0)$. Second, the level of disagreement. Next, Definition 1 formally describes an environment where O's opposition to the high action is of relatively low intensity, and Lemma 1 establishes the link between the intensity of opposition and the possibility that the FD exists.

**Definition 1** (Weak opposition)**.** We say that O's opposition to the high action is weak if both:
1. the ex ante probability that O is a zealot is low enough, i.e.,

$$F(0) < \frac{\Pr_S\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right)}{1 + \Pr_S\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right)}, \tag{12}$$

2. the disagreement is not too large,

$$q_0^O \geq q^*, \tag{13}$$

where $q^*$ is the unique value of $q_0^O$ such that $k_2 = k_3$, if a solution exists, or else $q^* = 0$.

**Lemma 1.** *FD is an equilibrium only if O's opposition to the high action is weak.*

S1's IC constraint is slacker than S2's when O's opposition is weak. Condition (12) of Assumption 1 requires that S1 is sufficiently confident that S2 will also report $H$, thereby inducing DM to choose $a = 1$. For S2 to also report $H$ after S1 reported $H$, two events have to occur. First, O should not cancel S1, otherwise S2 will always report $L$; and second, S2 should also observe a high signal. As the ex ante probability that O is a zealot increases, eventually S1 will be canceled with probability one. Hence, there is a point for $F(0)$ beyond which S1's incentive to experiment are irremediably compromised. Further, if $\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)$ is too small, then the probability that S2's signal will confirm S1's $H$ signal is too low for experimentation to be worthwhile. Next, we discuss the role of condition (13). This latter condition implies that the ex ante likelihood of S1 being canceled is not too high relative to the probability of S2 being canceled after first-period cancellation. To understand the connection between disagreement and each speaker's probability of being canceled, recall that O's equilibrium cutoff ($\hat{c}_{FD}$) is decreasing in his prior belief. Interestingly, an increase in O's cutoff for cancellation has opposite effects on S1's and S2's incentives. As O's prior decreases, the ex ante risk of S1 being canceled ($F(\hat{c}_{FD})$) increases, because more O types cancel. Consequently, the expected cost of S1's experimentation increases – all else equal making his IC constraint harder to satisfy. For S2, exactly the opposite happens. The intuition goes as follows. A higher cutoff means that more O types with a positive cost of canceling will punish S1 in the first period. Then, S2's posterior risk of being canceled conditional on first-period canceling ($F(0)/F(\hat{c}_{FD})$) decreases, because it is relatively less likely that S1 was canceled by a zealot, and only zealots will punish S2. Therefore, S2's cost of being truthful is lower – all else equal making it easier to satisfy his IC constraint. As long as O's prior remains above a certain $q^*$, the net benefit of truthfully disclosing $H$ is relatively greater for S1 than for S2, which gives the desired ordering of speakers' IC constraints.

## 3.4 Partial Deterrence Equilibrium

Observe that $k_2 > k_1$, meaning that the region of parameters where the FD exists (if it exists) does not overlap with the ND region. It follows that when $k \in (k_1, k_2)$ there is neither the FD nor the ND. In other words, a cost of being canceled $k \in (k_1, k_2)$ is too cheap to discourage truthtelling by S2 when he expects canceling with a lower likelihood (FD), but too costly to induce truthtelling by S2 when he expects canceling with a higher likelihood (ND). Conditional on O canceling S1, S2 has an incentive to report truthfully given O's cutoff $\hat{c}_{FD} > 0$. However, if both speakers are truthful in equilibrium, then only opponents with $c \leq 0$ would cancel. But then S2 would not be truthful, because after first-period cancellation he would face certain cancellation. This reasoning suggests that for $k \in (k_1, k_2)$ the equilibrium entails a mixed strategy by S2. We refer to this case as a "Partial Deterrence" (PD) equilibrium, to represent the idea that O's canceling in the first-period is only partially effective at blocking communication in the second period.

In particular, let us conjecture (and verify later) that, when S1 reports $H$ and is canceled by O, S2 truthfully reports $H$ with probability $\chi \in [0, 1]$. Let $\hat{c}_{PD}$ be O's equilibrium cutoff cost of canceling in the PD. S2's equilibrium mixing probability, denoted $\chi_{PD}$, must leave the cutoff O type indifferent between canceling and not canceling the S1 who reported $H$,

$$\begin{aligned} &\Pr_O\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right) \begin{bmatrix} \chi_{PD} V(1; q^O_{H,H}) \\ + \left(1 - \chi_{PD}\right) V(0; q^O_{H,H}) \end{bmatrix} + \Pr_O\left(\tilde{s}_2 = L | \tilde{s}_1 = H\right) V(0, q^O_{H,L}) - \hat{c}_{PD} \\ &= \Pr_O\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right) V(1; q^O_{H,H}) + \Pr_O\left(\tilde{s}_2 = L | \tilde{s}_1 = H\right) V(0; q^O_{H,L}), \end{aligned} \tag{14}$$

which simplifies to

$$\hat{c}_{PD} = (1 - \chi_{PD}) \hat{c}_{FD}. \tag{15}$$

O's indifference condition (14) is determined as follows. On the right-hand side of (15) we have O's payoff from not canceling S1. This expression is the same as the one on the right-hand side of (5): both in the FD and PD, S2 will be truthful after no first-period cancellation and, hence, DM takes the efficient action conditional on both speakers' signals. The left-hand side of (14), which

captures O's payoff from canceling S1, is different in the FD and PD. In comparison with the FD, here O's net benefit from canceling is lower, because S2 truthfully reveals $H$ with probability $\chi_{PD}$ despite first-period cancellation. Thus, we have $\hat{c}_{PD} < \hat{c}_{FD}$.

At the same time, S2 with signal $H$ is exactly indifferent between reporting $H$, thereby inducing $a = 1$, and reporting $L$, thereby inducing $a = 0$, if

$$V(1, q^S_{H,H}) - \Pr_S\left(\tilde{d}_2 = 1 | \tilde{d}_1 = 1\right) k = V(0, q^S_{H,H}), \tag{16}$$

where $\Pr_S\left(\tilde{d}_2 = 1 | \tilde{d}_1 = 1\right) = F(0)/F(\hat{c}_{PD})$ is S2's posterior probability of being canceled conditional on first-period cancellation. Rearranging (16), we obtain that, for S2 to be indifferent, O's cutoff $\hat{c}_{PD}$ must be

$$\hat{c}_{PD} = F^{-1}\left(\frac{F(0)\, k}{\Delta(q^S_{H,H})}\right). \tag{17}$$

Equation (17) gives O's cutoff in the PD explicitly as a function of exogenous parameters. Knowing $\hat{c}_{PD}$, we can recover S2's equilibrium mixing probability $\chi_{PD}$ from (15).

The PD is predicated on the first speaker having incentives to be truthful. Now S1 has an incentive to be truthful if

$$\begin{aligned} \Pr_S\left(\tilde{d}_1 = 1\right) & \begin{bmatrix} \Pr_S\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right) \chi_{PD} V(1, q^S_{H,H}) \\ + \Pr_S\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right) (1 - \chi_{PD})\, V(0, q^S_{H,H}) \\ + \Pr_S\left(\tilde{s}_2 = L | \tilde{s}_1 = H\right) V(0, q^S_{H,L}) - k \end{bmatrix} \\ + \Pr_S\left(\tilde{d}_1 = 0\right) & \begin{bmatrix} \Pr_S\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right) V(1, q^S_{H,H}) \\ + \Pr_S\left(\tilde{s}_2 = L | \tilde{s}_1 = H\right) V(0, q^S_{H,L}) \end{bmatrix} \geq V(0, q^S_H), \end{aligned} \tag{18}$$

where the probability of S1 being canceled is $\Pr_S\left(\tilde{d}_1 = 1\right) = F\left(\hat{c}_{PD}\right)$. S1's IC constraint in the PD, given by (18), is similar to his IC constraint in the FD, given by (9). The only difference between the two expressions is in S1's payoff from truthful disclosure of $H$ (left-hand side): if O cancels S1 and S2 observes the $H$ signal, in the PD there is a probability $\chi_{PD}$ that S2 is truthful

and DM chooses $a = 1$, whereas in the FD S2 will never be truthful and DM will always choose $a = 0$. Simplifying yields that S1 is willing to truthfully report $H$ if the cost of being canceled is less than some $k_3'$, which depends not only on O's cutoff $\hat{c}_{PD}$, but also on S2's mixing probability $\chi_{PD}$,

$$k \leq \frac{\left(1 - F\left(\hat{c}_{PD}\right) + \chi_{PD} F\left(\hat{c}_{PD}\right)\right) \Pr_S\left(\tilde{s}_2 = H | \tilde{s}_1 = H\right)}{F\left(\hat{c}_{PD}\right)} \Delta(q^S_{H,H}) \equiv k_3'. \tag{19}$$

S1's IC constraint in the PD is automatically satisfied whenever the FD interval $[k_2, k_3]$ is non-empty (i.e., $k_3' > k_3$). The reason is that if S1 is willing to be truthful in the FD, where first-period cancellation leads to a complete breakdown of communication in the second period, then clearly S1 wants to be truthful in the PD, where first-period cancellation is partially ineffective at deterring future information transmission.

## 3.5 No Information Transmission Equilibrium

The last equilibrium pattern is one in which it is too costly for both S1 and S2 to truthfully report $H$, which results in an equilibrium in which no information is transmitted. We denote this No Information Transmission by NIT. Given the discussion and analysis in the previous sections, it is easy to see that whenever $k$ is large enough the only equilibrium is NIT, in which both speakers never report $H$.

## 3.6 Equilibrium Taxonomy with Weak Opposition

We are now in a position to summarize the equilibrium taxonomy when O's opposition is weak. Note that all the conditions on the parameter values for each the equilibria (given a weak opposition) are mutually exclusive, and cover the entire support of the parameters. As such, we have identified a complete taxonomy of the equilibrium pattern, which we present in the following Proposition.

**Proposition 1.** *Suppose that O's opposition to the high action is weak. Then, one of the following four equilibrium configurations prevails. Which of the following equilibrium configuration prevails depends on the magnitude of the cost, $k$, that speakers incur when they are canceled:*

*(i) For $k \leq k_1$, there is no deterrence (ND): both speakers are truthful and only zealots cancel speakers.*

*(ii) For $k \in (k_1, k_2)$, there is partial deterrence of S2 (PD): in the first period, S1 is truthful and O cancels if his cost of canceling is less than $\hat{c}_{PD} > 0$; in the second period, S2 is always truthful after no first-period cancellation and truthfully reports $H$ only with probability $\chi_{PD}$ after first-period cancellation.*

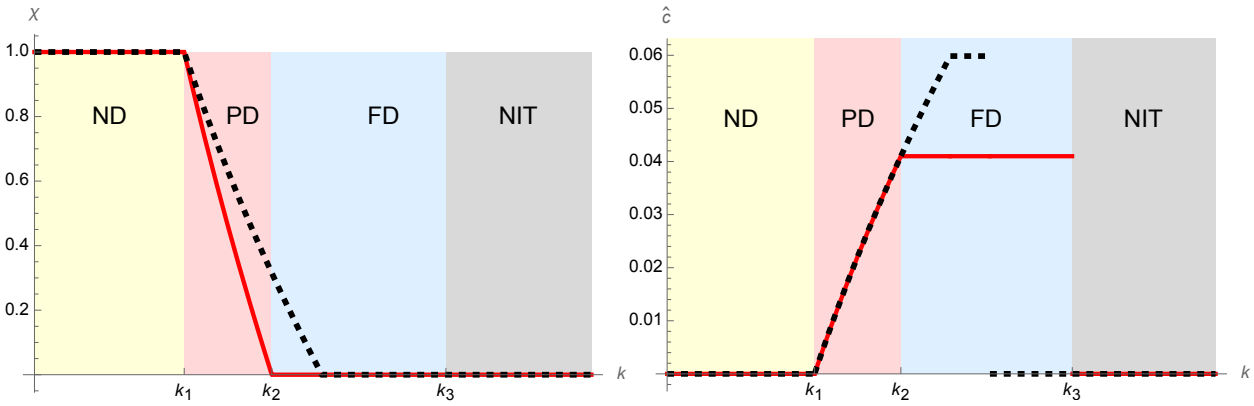
*(iii) For $k \in [k_2, k_3]$, there is full deterrence of S2 (FD): in the first period, S1 is truthful and O cancels if his cost of canceling is less than $\hat{c}_{FD} > \hat{c}_{PD}$; in the second period, S2 is truthful after no first-period cancellation but always reports $L$ after first-period cancellation.*

*(iv) For $k > k_3$, there is no information transmission (NIT): both speakers report $L$ regardless of their information.*

Figure 2 depicts the equilibrium taxonomy for the case of weak opposition. We can identify four regions for the cost of being canceled $k$ that speakers suffer when O cancels them. For low cost $k$ (yellow region), both speakers accept the risk of being canceled when they report $H$. No deterrence occurs and DM chooses the action knowing both speakers' signals. Therefore, informativeness of communication is maximal in this scenario. For moderately low cost (red region), S1 still reports truthfully, but after first-period cancellation S2 mixes between reporting $H$ truthfully, so as to induce DM to choose $a = 1$, and reporting $L$, so as to avoid potential cancellation by O. For moderately high costs (blue region), after first-period cancellation the threat of cancellation is too big for S2 and, consequently, he ceases to provide any information whatsoever. Only after no first-period cancellation will S2 be truthful, because in that case he expects no cancellation from a non-zealot O. Essentially, for intermediate cost (red and blue regions), S1 is experimenting to generate information about O's privately known cost of canceling – information which S2 exploits in the second period. By disclosing $H$, S1 forces O to make a move and, if O does not cancel, then S2 can infer that O is not a zealot. Even though cancellation is personally costly to S1, he has an incentive to experiment because S2 might have information that is pivotal for DM's choice. Finally, when cancellation is too costly for speakers (gray region), the expected cost of experimenting outweighs its benefit, and even S1 becomes unwilling to disclose $H$. Equilibrium informativeness is

minimal here, because DM chooses the action based only on prior information.

To conclude the analysis of equilibria when O's opposition is weak, observe that the condition $q_0^O \geq q^*$ guarantees continuity of S2's equilibrium strategy as a function of $k$. Indeed, when $k = k_1$ we have $\chi_{PD} = 1$ and $\hat{c}_{PD} = 0$, as in the ND; and when $k = k_2$ we have $\chi_{PD} = 0$ and $\hat{c}_{PD} = \hat{c}_{FD}$, as in the FD.




(a) (b)

Figure 2: Equilibrium strategies and cost of being canceled: case of weak opposition. The figures depict the four possible equilibrium configurations depending on the cost $k$ that speakers incur when O cancels them. In panel (a), the thick red graph is the probability $\chi$ that S2 truthfully reports $H$, after first-period cancellation, as a function of $k$. In panel (b), the thick red graph is O's cancellation cutoff $\hat{c}$ as a function of $k$. The dashed black graphs show the effects of an increase in disagreement (i.e., a decrease in O's prior). The parameters are: $q_0^{DM} = q_0^S = .2$, $\alpha = .75$, $q_0^O = .08$ ($q_0^O = .073$) for the thick red (dashed black) graphs, and $\tilde{c} \sim \mathcal{N}(.2, .15^2)$.

## 3.7 Strong Opposition

We now turn to situations where O's opposition to the high action is strong. We refer to strong opposition as parameter configurations such that the ex ante probability that O is a zealot, or disagreement, are relatively high. Formally, the region of parameters we hereby define as strong opposition is complementary to the region we identified as weak opposition.

**Definition 2** (Strong opposition)**.** We say that O's opposition to the high action is strong if condition (12) or (13) is violated (or both).

When O's opposition is strong, the FD interval is empty (see Lemma 1). Intuitively, when speakers face an O who is strongly opposed to the high action, S1 does not have enough incentives

to experiment whether O is a zealot: S1's probability of being canceled is too large compared to the benefit of enabling S2 to be truthful only after no first-period cancellation. As a consequence, the FD cannot exist. However, S1 might still be incentivized to truthfully report $H$ if S2 is truthful with some probability even after first-period cancellation, as long as S1's risk of being canceled is still within reasonable bounds.

**Proposition 2.** *Suppose O's opposition to the high action is strong. Then, one of the following three equilibrium configurations prevails. Which equilibrium configuration prevails depends on the magnitude of the cost, $k$, that speakers incur when they are canceled.*

*(a) If the ex ante probability that O is a zealot $F(0) < \Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)$, then:*

*(i) For $k \leq k_1$, there is no deterrence (ND).*

*(ii) For $k \in (k_1, k_4]$, where $k_4$ is the unique solution to*

$$k = k_3'|_{\hat{c}_{PD}=F^{-1}\left(\frac{F(0)k}{\Delta(q^S_{H,H})}\right)},$$

*there is partial deterrence (PD).*

*(iii) For $k > k_4$, there is no information transmission (NIT).*

*(b) If $F(0) \geq \Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)$, then:*

*(i) For $k \leq k_1$, there is ND.*

*(ii) For $k > k_1$, there is NIT.*

The proposition states that, under strong opposition by O, the equilibrium taxonomy features a discontinuity. Consider panels (a) and (b) of Figure 3. As the speakers' cost of being canceled $k$ grows, players' strategies transition smoothly from the ND (yellow region) to the PD (red region). But as $k$ keeps growing, at some point S1's IC constraint cannot be satisfied any longer, and so the entire information transmission process breaks down (gray region). Though S2 would be willing to convey some information if S1 reported $H$, S1 always reports $L$ because the risk of being canceled is too big. In panels (c) and (d) of Figure 3, S1's incentive problem is even more severe. In that case, the ex ante probability that S1 faces a zealot O is so large that, even if S2 were to be truthful in

spite of O's cancellation, S1 could not be induced to truthfully report $H$. Consequently, as $k$ grows we transition abruptly from the ND (yellow region) to the NIT (gray region). Overall, Proposition 2 demonstrates that small changes in some parameter can have a drastic impact on the amount of information transmitted in equilibrium.

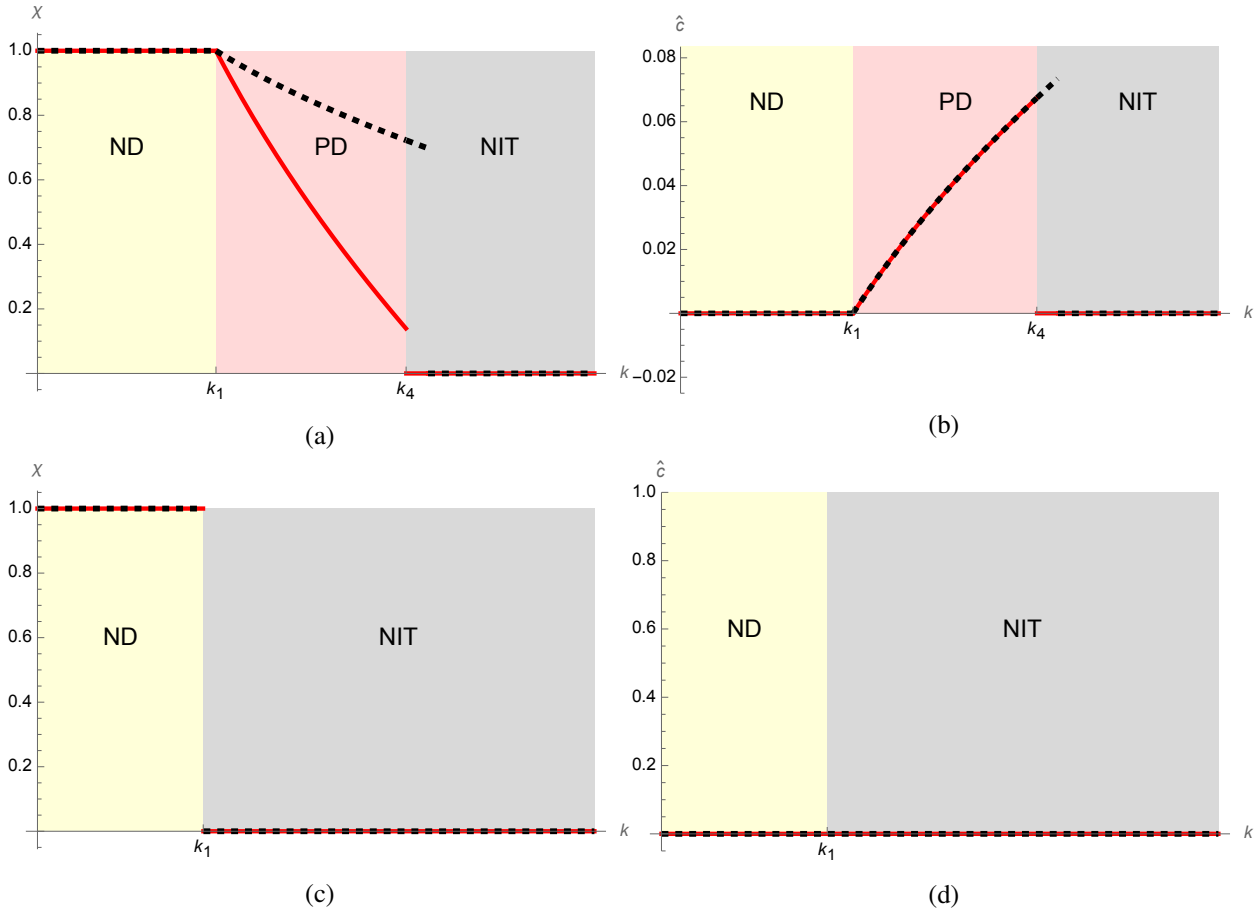



Figure 3: Equilibrium strategies and cost of being canceled: case of strong opposition. The figures depict the three possible equilibrium configurations depending on the cost $k$ that speakers incur when O cancels them. In panels (a) and (c), the thick red graph is the probability $\chi$ that S2 truthfully reports $H$, after first-period cancellation, as a function of $k$. In panels (b) and (d), the thick red graph is O's cancellation cutoff $\hat{c}$ as a function of $k$. The black dashed graphs show the effects of an increase in disagreement (i.e., a decrease in O's prior). In panels (a)-(b), the ex ante probability that O is a zealot $F(0)$ is relatively low (part (a) of Proposition 2), whereas in panels (b)-(c) $F(0)$ is relatively high (part (b) of Proposition 2). The parameters are: $q_0^{DM} = q_0^S = .2$, $\alpha = .75$, $q_0^O = .065$ ($q_0^O = .002$) for the thick red (dashed black) graphs, $\tilde{c} \sim \mathcal{N}(.2, .15^2)$ in panels (a)-(b), and $\tilde{c} \sim \mathcal{N}(0, .15^2)$ in panels (c)-(d).

# 4 Comparative Statics

In this section, we study how the equilibrium informativeness varies with the parameters of the model. Recall that we define informativeness as the probability that both speakers report truthfully conditional on both of them observing a high signal. We denote this probability by $\tau(H, H)$.

Informativeness is the greatest in the ND, because both speakers truthfully disclose their signals and, therefore, DM chooses the action under symmetric information. Vice versa, informativeness is the smallest in the NIT, because speakers' reports do not convey any information whatsoever and, hence, DM choose the action based on the prior.

In the FD and PD, speakers' reports are neither fully informative nor fully uninformative. In these equilibrium configurations, informativeness is jointly determined by O's cutoff for cancellation ($\hat{c}$) and the probability that S2 is truthful after first-period cancellation ($\chi$). As O's cutoff increases, O cancels S1 more often. When S1 is canceled, further communication by S2 is deterred. All else equal, the effect is to reduce informativeness. By contrast, informativeness increases as the probability that S2 is truthful, in spite of first-period cancellation, increases. In addition, changes in the model parameters have an impact on equilibrium informativeness to the extent that they alter the boundaries of the different equilibrium regions.

**Level of disagreement.** We begin the comparative statics analysis by taking into consideration the effects of changes in the level of disagreement. Fixing speakers' and DM's priors, more disagreement is captured by a lower O's prior ($q_0^O$). Proposition 3 below demonstrates that the effect of increases in disagreement on informativeness are non-monotonic.

**Proposition 3.** *(i) Suppose the ex ante probability that O is a zealot* $F(0) < \frac{\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}{1+\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}$. *For each value* $k$ *of speakers' cost of being canceled, there exists a cutoff* $q_k^*$ *such that equilibrium informativeness is weakly decreasing (increasing) in* $q_0^O$ *for* $q_0^O < q_k^*$ *(*$q_0^O > q_k^*$*).*

*(ii) Suppose* $F(0) \geq \frac{\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}{1+\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}$. *Equilibrium informativeness is weakly decreasing in* $q_0^O$.

Why does a larger disagreement (i.e., a lower $q_0^O$) sometimes stimulate information transmis-

sion? There are two effects that lead to this result. One such effect is present in the FD region: S2's IC constraint relaxes as disagreement gets larger. As a consequence, in Figure 2 the equilibrium configuration may shift from FD to PD. In the FD region, O is more likely to cancel S1 when disagreement is larger. But then, first-period cancellation becomes less credible as a threat for S2: it becomes relatively more likely that cancellation came from an O with positive cost of canceling, who would not cancel S2 in the second period, when cancellation has no deterrence role. This decrease in S2's posterior probability of being canceled can incentivize S2 to be truthful with some probability, thus increasing informativeness.

The second positive effect on informativeness is present in the PD region: as disagreement increases, S2 becomes more likely to be truthful after first-period cancellation. This can be seen in the PD region of both Figures 2 and 3(a)-(b), where the black dashed graph (larger disagreement) is above the red thick graph (smaller disagreement). To understand this effect, note the following. On the one hand, O can deter S2's communication by canceling S1. On the other hand, O is discouraged from canceling S1 when S2 is more often truthful despite first-period cancellation. All else equal, O cancels more aggressively in the first period when disagreement is larger. However, in the PD region, S2 responds to an increase in disagreement by being truthful more often, thereby destroying O's additional incentives to cancel S1. As a result, the probability that O cancels S1 is unchanged. Overall, S1 is canceled with the same probability and S2 is more likely to be truthful, which increases informativeness. For the same reasons, S1's IC constraint relaxes, thus shrinking the NIT region in Figure 3(a)-(b) and also contributing to higher informativeness.

A larger disagreement also has two negative effects on informativeness. First, in the FD region, it increases the mass of O types that cancel in the first period. Because in the FD first-period cancellation is fully effective at deterring future communication, this translates into lower informativeness. The second negative effect is that, as disagreement becomes larger, the equilibrium configuration may shift from FD to NIT (see Figure 2(b)). The reason is that once O starts canceling too often in the first period, S1 cannot be incentivized any longer to disclose truthfully, thus resulting in lower informativeness.

Finally, observe that all the results in Proposition 3 have been stated in terms of weak monotonicity. This is because when speakers' cost of being canceled $k$ is very small, the ND can be sustained irrespective of the level of disagreement, which then has no effect on informativeness. Similarly, when $k$ is very high, only the NIT can be sustained for any level of disagreement, which once again has no effect on informativeness. It is for intermediate values of $k$ that the effects are strict. For such intermediate values, Figure 4 illustrates the positive and negative effects of disagreement on equilibrium informativeness. The ex ante probability that O is a zealot is relatively low in panels (a)-(c), corresponding to Proposition 3(i), and relatively high in panel (d), corresponding to Proposition 3(ii). Whenever the PD region exists, in that region informativeness increases in disagreement (i.e., decreases in $q_0^O$), because S2 reports truthfully more often. Oppositely, whenever the FD region exists, in that region informativeness decreases in disagreement (i.e., increases in $q_0^O$), because more O types cancel S1, thereby deterring communication by S2. Further, a larger disagreement increases informativeness when it shifts the equilibrium from NIT to PD (because S1's IC constraint relaxes) and decreases informativeness when it shifts the equilibrium from FD to NIT (because S1's IC constraint can no longer be satisfied).

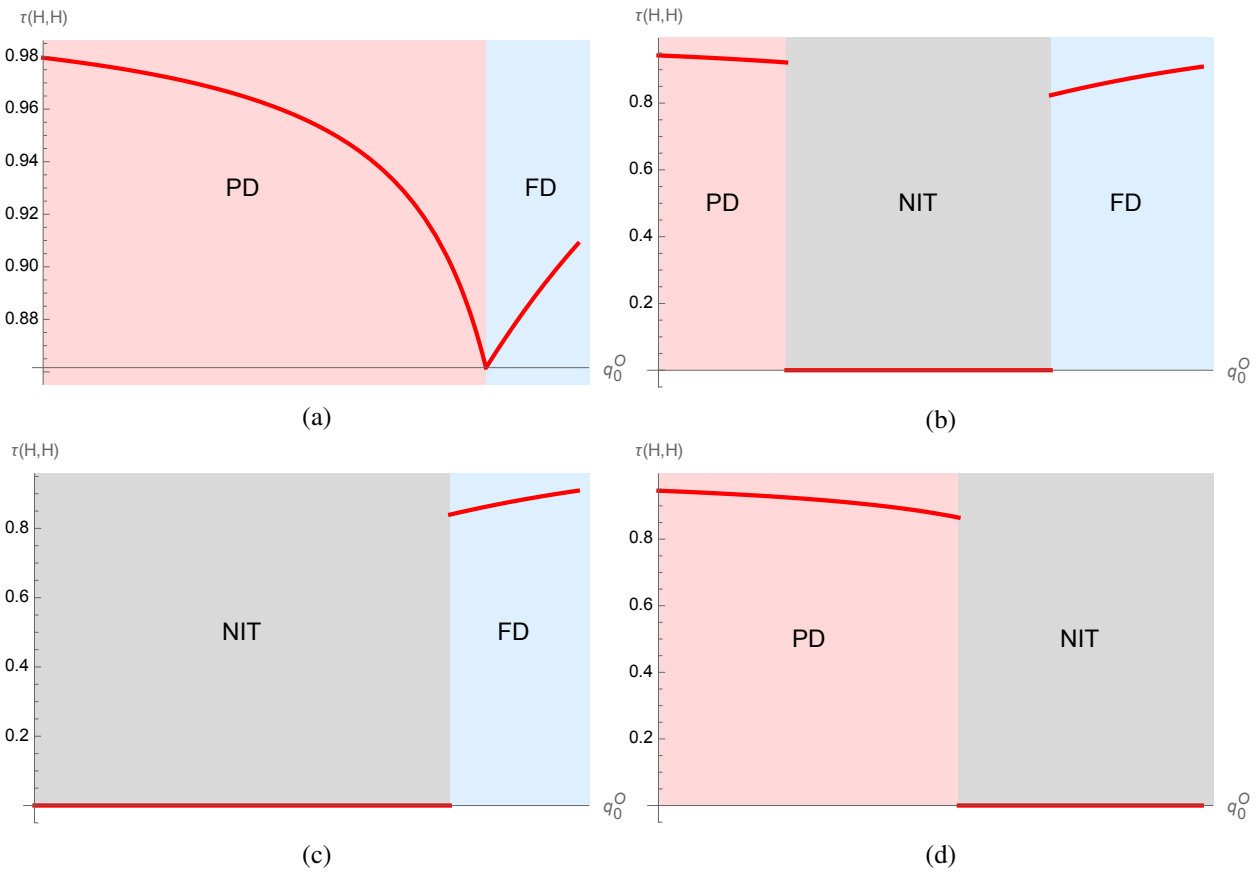



Figure 4: Information transmission and the level of disagreement. The figures depict equilibrium informativeness $\tau(H,H)$ (i.e., the probability that both speakers report truthfully conditional on both of them observing a high signal) as a function of O's prior $q_0^O$. The ex ante probability that O is a zealot is relatively low in panels (a)-(c), corresponding to Proposition 3(i), and relatively high in panel (d), corresponding to Proposition 3(ii). The parameters are: $q_0^{DM} = q_0^S = .2$, $\alpha = .75$, $k = 0.584$ in panel (a), $k = 0.833$ in panel (b), $k = 0.936$ in panel (c), $k = 0.415$ in panel (d), $\tilde{c} \sim \mathcal{N}(.2, .15^2)$ in panels (a)-(c), and $\tilde{c} \sim \mathcal{N}(.2, .5^2)$ in panel (d).

**Speakers' cost of being canceled.** While the effect of disagreement on informativeness is ambiguous, the effect of speakers' cost of being canceled ($k$) is always monotonic.

**Proposition 4.** *Equilibrium informativeness is weakly decreasing in speakers' cost of being canceled $k$.*

The result of Proposition 4 is as expected. The reason why here the comparative statics is unambiguous, and the difference with the case of disagreement, is that $k$ increases the relative cost of being truthful for both speakers. This can only lead to a reduction in informativeness. By

contrast, disagreement affects the two speakers' incentives in opposite ways: a larger disagreement discourages S1 from disclosing $H$ (because of greater ex ante probability of being canceled), but encourages S2 to disclose $H$ (because of lower posterior probability of being canceled).

# 5 Conclusion

We study learning as a sequential and collective random process, wherein speakers face the risk of bearing the cost of being cancelled. We demonstrate that cancellations cannot arise in equilibrium unless there is a positive probability that the opponent enjoys a personal benefit from cancelling speakers who dissent from the status quo. We also demonstrate that the risk of cancellation can be effective at disrupting information transmission. Such costs might be more substantial when institutional protection for dissent is weaker.

In our model, we made a number of simplifying assumptions. First we assume that the beliefs of speakers and DM are relatively similar, whereas the Opponent holds beliefs that are substantially different. Equivalently, we assumed relatively low disagreement between the senders and DM regarding the optimal action, and relatively high disagreement between them and Opponent. Our analysis does not crucially depend on that, but if the senders were sufficiently in disagreement with DM, this would affect their ability and incentive to transmit information to DM. We assume for simplicity of disposition that the Opponent's beliefs are such that he prefers one particular action. However, the results would be symmetric if the disagreement were to be in the opposite direction.

Further, we consider a stylized model with a single opponent with unknown type. In reality, there are potentially multiple opponents who face a coordination problem when they decide to attack a speaker. In particular, opponents do not know how many other opponents will end up attacking/cancelling the speaker and whether the attack will be "effective." This aspect could be added to the model, although it would complicate it significantly without necessarily qualitatively changing the main results.

Herbert Marcusse (Marcuse, 1965) argued that tolerance and free speech confer benefits on

society only under a special condition that almost never exists: absolute equality. Inspired by these ideas, the cancel culture's worldview maintains that speech per se can be unfair, dangerous, and harmful, and that it is legitimate to suppress it according to the preferences of groups that activists believe to be oppressed. At the same time, as asserted by the Harper letter, an argument can be made that "The restriction of debate, whether by a repressive government or an intolerant society, invariably hurts those who lack power and makes everyone less capable of democratic participation. The way to defeat bad ideas is by exposure, argument, and persuasion, not by trying to silence or wish them away." While our model does not address this question, we hope that future analytical research will be able to shed light on it.

# References


Acemoglu, D. and A. Wolitzky (2023). Mistrust, misperception, and misunderstanding: Imperfect information and conflict dynamics. Technical report, National Bureau of Economic Research.

Aghion, P., P. Bolton, C. Harris, and B. Jullien (1991). Optimal learning by experimentation. The review of economic studies 58(4), 621–654.

Antic, N., A. Chakraborty, and R. Harbaugh (2020). Subversive conversations. working paper.

Ben-Porath, E., E. Dekel, and B. L. Lipman (2018). Disclosure and choice. The Review of Economic Studies 85(3), 1471–1501.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1992). A theory of fads, fashion, custom, and cultural change as informational cascades. Journal of political Economy 100(5), 992–1026.

Bikhchandani, S., D. Hirshleifer, and I. Welch (1998). Learning from the behavior of others: Conformity, fads, and informational cascades. Journal of economic perspectives 12(3), 151–170.

Bikhchandani, S., D. Hirshleifer, and I. Welch (2016). Information cascades. In The New Palgrave Dictionary of Economics, pp. 1–9. London: Palgrave Macmillan UK.

Che, Y.-K. and N. Kartik (2009). Opinions as incentives. Journal of Political Economy 117(5), 815–860.

Chen, Y., K. Du, P. Stocken, and Z. Wang (2024). Peer learning, enforcement, and reputation. working paper.

Crawford, V. P. and J. Sobel (1982). Strategic information transmission. Econometrica 50(6), 1431–1451.

Goldstein, I. and L. Yang (2017). Information disclosure in financial markets. Annual Review of Financial Economics 9, 101–125.

Hayek, F. A. (1945). The use of knowledge in society. The American economic review 35(4), 519–530.

Keller, G. and S. Rady (1999). Optimal experimentation in a changing environment. The review of economic studies 66(3), 475–507.

Keller, G., S. Rady, and M. Cripps (2005). Strategic experimentation with exponential bandits. Econometrica 73(1), 39–68.

Kreps, D. M. and R. Wilson (1982). Reputation and imperfect information. Journal of economic theory 27(2), 253–279.

Loury, G. C. (1994). Self-censorship in public discourse: A theory of "political correctness" and related phenomena. Rationality and Society 6(4), 428–461.

Lowery, R. and C. M. Carvalho (2021). How critical theory fundamentally challenges traditional inquiry in social science. Available at SSRN 3941223.

Marcuse, H. (1965). Repressive tolerance.

Morris, S. (1995). The common prior assumption in economic theory. Economics & Philosophy 11(2), 227–253.

Morris, S. (2001). Political correctness. Journal of political Economy 109(2), 231–265.

Morris, S. and H. S. Shin (2003). Global games: Theory and applications. In Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress, Volume 1, pp. 56–114. Cambridge University Press.

Ottaviani, M. and P. N. Sørensen (2006). Professional advice. Journal of Economic Theory 126(1), 120–142.

Ottaviani, M. and P. N. Sorensen (2006). Reputational cheap talk. The RAND Journal of Economics 37(1), 155–175.

Selten, R. (1978). The chain store paradox. Theory and decision 9(2), 127–159.

Sims, C. A. (2003). Implications of rational inattention. Journal of monetary Economics 50(3), 665–690.

# Appendix

# A Proofs of Section 3

## A.1 Ruling out equilibria that are not NIT, ND, FD, or PD

To begin with, we provide a series of lemma to establish that the only possible equilibrium configurations are those identified in the main text: NIT, ND, FD, and PD. Specifically, the proof involves the following steps.

- Lemma A.1 pins down the action of an S1 who observed $s_1 = L$ (he truthfully reports low) and the continuation game after a low report by S1 (S2's report is uninformative and the DM chooses the low action).
- Next, there are two possibilities regarding the action of an S1 who observed $s_1 = H$.

  - If such an S1 reports low with probability one, then there is no communication at all in equilibrium. This is the NIT.

  - Instead, suppose that such an S1 truthfully reports high with positive probability. Because S1 truthfully reports $s_1 = L$, then following $r_1 = H$ players know that S1's signal was high. The action that the DM eventually takes will depend on S2's report, which itself depends on O's cancellation decision in the first period.
- Lemma A.2 shows that, following $r_1 = H$, the DM chooses the low action when $r_2 = L$. Further, if there was no cancellation in the first period, then the DM chooses the high action when $r_2 = H$.
- Lemma A.3 shows that S2 truthfully reports $s_2 = L$, truthfully reports $s_2 = H$ after no first-period cancellation, but may misreport $s_2 = H$ after first-period cancellation. If after first-period cancellation:

  - S2 truthfully reports $s_2 = H$, then we are in the ND;

  - S2 never truthfully reports $s_2 = H$, then we are in the FD;

  - S2 truthfully reports with some probability $s_2 = H$, then we are in the PD.
- Last, Lemma A.4 shows that S1 does not mix.

**Lemma A.1.** *Let $\hat{q}_{r_1} \equiv \Pr_{DM}(\tilde{\theta} = 1 | \tilde{r}_1 = r_1)$ denote the DM's posterior belief about the state of nature given that S1 reported $r_1$.*

*(i) In continuation game after S1 reports $r_1 = L$, the DM chooses $a = 0$ and S2's report is uninformative.*

*(ii) In equilibrium, S1 truthfully reports the low signal with probability one.*

*(iii) If in equilibrium S1 truthfully reports the high signal with positive probability, then the DM's posterior given the $r_1 = H$ is $\hat{q}_H = q_H$.*

*Proof of Lemma A.1. Proof of part (i).* The DM's posterior belief given a low report by S1 is such that $\hat{q}_L \leq q_0$. It follows that, in the second period, the DM's posterior can be at most

$$\Pr_{DM}(\tilde{\theta} = 1 | \tilde{r}_1 = L, \tilde{s}_2 = H) = \frac{\hat{q}_L \alpha}{\hat{q}_L \alpha + (1 - \hat{q}_L)(1 - \alpha)} \leq \frac{q_0 \alpha}{q_0 \alpha + (1 - q_0)(1 - \alpha)} = q_H,$$

and we assumed that $q_H$ is not high enough for the DM to take the high action. Next, given that in the continuation game the DM chooses the low action, S2 has no incentive to report $r_2 = H$: a high report by S2 does not affect the DM's action and, with probability $F(0)$, is punished by the zealot.

*Proof of part (ii).* We assumed that if S1 observes the low signal, then S1 prefers the low action regardless of what S2's signal turns out to be. As shown in part (i) of this lemma, reporting $r_1 = L$ ensures that the DM takes the low action. S1's payoff from reporting instead $r_1 = H$ would be strictly lower, because S1 expects to be punished with positive probability and the DM might eventually choose the high action.

*Proof of part (iii).* By part (ii) of this lemma, a high report in the first period can come only from a speaker that observed the high signal. □

**Lemma A.2.** *Let $a(r_1, d_1, r_2)$ denote the DM's action as a function of the history (we omit $d_2$ because it does not affect the DM's choice). Suppose that, in equilibrium, S1 truthfully reports the high signal with positive probability.*

*(i) In the continuation game after S1's report $r_1 = H$ and any first-period cancellation decision $d_1 \in \{0, 1\}$, we have $a(H, d_1, L) = 0$.*

*(ii) In the continuation game after $r_1 = H$ and $d_1 = 0$, we have $a(H, 0, H) = 1$.*

*Proof of Lemma A.2. Proof of part (i).* Let $\hat{q}_{r_1, d_1, r_2} \equiv \Pr(\tilde{\theta} = 1 | \tilde{r}_1 = r_1, \tilde{d}_1 = d_1, \tilde{r}_2 = r_2)$ denote the DM's posterior belief about the state of nature given that S1 reported $r_1$, O's cancellation decision was $d_1$, and S2 reported $r_2$. We have $\hat{q}_{H, d_1, L} \leq \mathrm{E}_{DM}\left(\hat{q}_{r_1, d_1, \tilde{r}_2} | \tilde{r}_1 = H, \tilde{d}_1 = d_1\right) = q_H$, where the inequality follows from the law of iterated expectations and the equality from Lemma A.1(iii). We assumed that the posterior $q_H$ is not high enough for the DM to take the high action, hence $a(H, d_1, L) = 0$.

*Proof of part (ii).* By contradiction, suppose that $a(H,0,H)=0$. By part (i) of this lemma, we have $a(H,0,L)=a(H,0,H)=0$. We distinguish two cases for the continuation game after $d_1=1$: the case where $a(H,1,L)=a(H,1,H)=0$ and the case where $0=a(H,1,L)<a(H,1,H)=1$ (S2's report $r_2=H$ may or may not be on the equilibrium path).

**Case** $a(H,1,L)=a(H,1,H)=0$**.** If this were the case, it would mean that the DM chooses the low action irrespective of whether S1 reports high or low. But then S1 would be strictly better off by reporting low, to avoid the expected cost of being canceled – this contradicts the assumption of this lemma that S1 reports the high signal truthfully with positive probability.

**Case** $0=a(H,1,L)<a(H,1,H)=1$**.** If S2 reported $r_2=L$ with probability one, then again the DM would choose the low action irrespective of whether S1 reports high or low, and we reach the same contradiction as in the previous case. If S2 reported $r_2=H$ with positive probability, then the DM would choose the low action after no first-period cancellation but sometimes choose the high action after first-period cancellation. It follows that all O types with a positive cost of canceling are better off not canceling (to ensure $a=0$). After first-period cancellation, an S2 who observes $s_2=L$ has a strict incentive to report low (to avoid certain cancellation by O types with $c<0$). Because $\hat{c}=0$, for S2 to have any incentive to truthfully disclose $s_2=H$, we must have

$$k\leq\Delta(q^S_{H,H}) \tag{A.1}$$

(see 2). Taking into account the strategies in the continuation game, for S1 to truthfully disclose $s_1=H$ we must have

$$\begin{aligned}&F(0)\begin{bmatrix}\Pr_S\left(\tilde{s}_2=H|\tilde{s}_1=H\right)V(1,q^S_{H,H})\\+\Pr_S\left(\tilde{s}_2=L|\tilde{s}_1=H\right)V(0,q^S_{H,L})-k\end{bmatrix}+(1-F(0))V(0,q^S_H)\\&\geq V(0,q^S_H)\iff k\leq\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)\Delta(q^S_{H,H}).\end{aligned} \tag{A.2}$$

Combining (A.1) and (A.2), we find that, for such an equilibrium to exist, we must have $k\leq\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)\Delta(q^S_{H,H})$. Therefore, if such an equilibrium exists, then also the ND exists. Because the ND is more informative, we select the ND instead of this equilibrium. □

**Lemma A.3.** *Suppose that, in equilibrium, S1 truthfully reports the high signal with positive probability.*

*(i) In the continuation game after S1's report $r_1=H$ and any first-period cancellation decision $d_1\in\{0,1\}$, S2 truthfully reports $s_2=L$.*

*(ii) In the continuation game after $r_1=H$ and $d_1=0$, S2 truthfully reports $s_2=H$.*

*Proof of Lemma A.3. Proof of part (i).* An S2 who observed $s_2=L$ prefers the DM to take the low action, and reporting $r_2=L$ ensures $a=0$ (Lemma A.2(i)) and no cancellation.

*Proof of part (ii).* No first-period cancellation reveals that O has a positive cost of canceling. Hence, in the second period S2 can report high without being canceled. Further, from Lemma A.2(ii) we know that S2 can induce the high action by reporting high. Because an S2 who observed $r_2 = H$ prefers the DM to take the high action, in equilibrium such an S2 truthfully reports $s_2 = H$. □

**Lemma A.4.** *In equilibrium, generically S1 either reports $r_1 = L$ with probability one or truthfully reports with probability one.*

*Proof of Lemma A.4.* We know from Lemma A.1 that an S1 who observed $s_1 = L$ truthfully reports low. So, if the equilibrium involves any mixing by S1 at all, it must be by the S1 who observed $s_1 = H$. Note that mixing by such an S1 would not affect the set of possible continuation games: after $r_1 = H$, the DM's posterior is $\hat{q}_H = q_H$ (the same as if S1 truthfully reported S1 with probability one); and after $r_1 = L$, the DM chooses $a = 0$. Lemmas A.2 and A.3 rule out any equilibrium with information transmission other than the ND, PD, and FD. Generically, S1's IC constraint in these equilibria holds with a strict inequality, hence no mixing. □

## A.2 Proofs of NIT, ND, FD, and PD equilibria

*Proof of Lemma 1.* We have $\partial(k_3 - k_2)/\partial\hat{c}_{FD} < 0$. Hence, if

$$(k_3 - k_2)|_{\hat{c}_{FD}=0} < 0 \iff F(0) > \frac{\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)}{1 + \Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)},$$

then the interval is empty. The complementary inequality gives us condition (12). If (12) is satisfied, then $(k_3 - k_2)|_{\hat{c}_{FD}=0} > 0$. Further, we have $\lim_{\hat{c}_{FD}\to\infty}(k_3 - k_2) < 0$. By continuity, there must exist a $\hat{c}^* \in (0, \infty)$ such that $k_3 - k_2 = 0$. Because $k_3 - k_2$ is decreasing in $\hat{c}_{FD}$, such a $\hat{c}^*$ is unique. Hence, for the interval to be non-empty we need $\hat{c}_{FD} \leq \hat{c}^*$. To obtain condition (13) we simply restate the inequality $\hat{c}_{FD} \leq \hat{c}^*$ in terms of the parameter $q_0^O$. Let $\bar{q}_0^O \equiv \frac{(1-\alpha)^2}{1-2\alpha(1-\alpha)}$ denote the upper bound on the prior $q_0^O$ implied by Assumption 2. Observe that $\hat{c}_{FD}$ is decreasing in $q_0^O$ (see Lemma B.5(i)) and satisfies $\hat{c}_{FD}|_{q_0^O=\bar{q}_0^O} = 0$. Therefore, there exists a $q^* \in [0, \bar{q}_0^O)$ such that $\hat{c}_{FD} \leq \hat{c}^*$ for all $q_0^O \geq q^*$. If $\hat{c}_{FD}|_{q_0^O=0} > \hat{c}^*$, then $q^* \in (0, \bar{q}_0^O)$, whereas if $\hat{c}_{FD}|_{q_0^O=0} \leq \hat{c}^*$, then $q^* = 0$. □

*Proof of Proposition 1.* The steps to construct the equilibrium, for each of the equilibrium configurations ND, PD, FD, and NIT, have been described in the main text. Lemma 1 showed that the assumption of weak opposition implies $k_2 < k_3$. The inequality $k_1 < k_2$ follows from inspection of (4) and (8). To complete the proof of this proposition, there remains to show that for

$k \in (k_1, k_2)$ the mixing probability $\chi_{PD}$ implied by (15) is indeed between zero and one. Solving (15) explicitly for $\chi_{PD}$, and plugging in the equilibrium value of $\hat{c}_{PD}$ from (17), we obtain

$$\chi_{PD} = 1 - \frac{F^{-1}\left(\frac{F(0)k}{\Delta(q^S_{H,H})}\right)}{\hat{c}_{FD}}, \tag{A.3}$$

whence

$$\begin{aligned} \chi_{PD} > 0 &\iff k < k_2, \\ \chi_{PD} < 1 &\iff k > k_1, \end{aligned} \tag{A.4}$$

where $k_1 = \min\left\{\frac{\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)}{F(0)}\Delta(q^S_{H,H}), \Delta(q^S_{H,H})\right\} = \Delta(q^S_{H,H})$ follows from weak opposition. Therefore, (A.4) states that $\chi_{PD} \in (0,1)$ if and only if $k \in (k_1, k_2)$, which is what we wanted to show. □

*Proof of Proposition 2. Proof of part (a).* The FD is impossible when O's opposition is strong, because in this case $k_3 < k_2$, as shown in the proof of Lemma 1. Sub-parts (i) and (iii) have been derived in the main text. Therefore, here we only need to prove sub-part (ii). According to (19), in the PD the IC constraint of S1 requires $k \leq k_3'$. Plugging into the expression for $k_3'$ the equilibrium values of $\chi_{PD}$ (equation (A.3)), $\hat{c}_{PD}$ (equation (17), and $\hat{c}_{FD}$ (equation 6) gives $k_3'$ as a function of $k$ and $q_0^O$, $k_3'(k, q_0^O)$. Define the function $\Gamma(k, q_0^O) \equiv k - k_3'(k, q_0^O)$. With this notation, S1's IC constraint in the PD is satisfied if $\Gamma(k, q_0^O) \leq 0$. Under the assumption for $F(0)$ of part (a) of the proposition, $k_1 = \Delta(q^S_{H,H})$. Note the following properties of $\Gamma(k, q_0^O)$:

$$\begin{aligned} \Gamma(k_1, q_0^O) &= k_1 - k_3'|_{\chi_{PD}=1, \hat{c}_{PD}=0} \\ &= \Delta(q^S_{H,H}) - \frac{\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)}{F(0)}\Delta(q^S_{H,H}) < 0, \\ \Gamma(k_2, q_0^O) &= k_2 - k_3'|_{\chi_{PD}=0, \hat{c}_{PD}=\hat{c}_{FD}} = k_2 - k_3 > 0, \end{aligned} \tag{A.5}$$

where the inequality for $\Gamma(k_1, q_0^O)$ follows from the assumption on $F(0)$ for part (a) of the proposition, and the inequality for $\Gamma(k_2, q_0^O)$ from strong opposition. By continuity, a solution $k_4$ exists in the interval $(k_1, k_2)$. The solution is unique because

$$\frac{d\Gamma(k, q_0^O)}{dk} = 1 - \left(\underbrace{\frac{\partial k_3'}{\partial \chi_{PD}}}_{>0}\underbrace{\frac{\partial \chi_{PD}}{\partial k}}_{<0} + \underbrace{\frac{\partial k_3'}{\partial \hat{c}_{PD}}}_{<0}\underbrace{\frac{\partial \hat{c}_{PD}}{\partial k}}_{>0}\right) > 0. \tag{A.6}$$

We conclude that S1's IC constraint for the PD is satisfied for $k \in (k_1, k_4)$ and violated for $k \in (k_4, k_2)$. We need not consider values for $k$ outside of the inteval $(k_1, k_2)$, because S1's mixing

probability would not be between zero and one (see (A.4)). Thus, the PD exists if and only if $k \in (k_1, k_4)$.

*Proof of part (b).* Both sub-parts have been derived in the main text. Also, and as in part (a), the FD does not exist when O's opposition is strong. There remains to prove that neither the PD exists. Under the assumption for $F(0)$ of part (b) of the proposition, $k_1 \leq \Delta(q^S_{H,H})$. We already know that the PD does not exist for $k \in (0, \Delta(q^S_{H,H})) \cup (k_2, \infty)$, because S1's mixing probability would not be between zero and one (see (A.4)). As to $k \in (\Delta(q^S_{H,H}), k_2)$, note that

$$\begin{aligned}\Gamma(\Delta(q^S_{H,H}), q^O_{H,H}) &= \Delta(q^S_{H,H}) - k_3'|_{\chi_{PD}=1, \hat{c}_{PD}=0} \\ &= \Delta(q^S_{H,H}) - \frac{\Pr_S(\tilde{s}_2 = H | \tilde{s}_1 = H)}{F(0)} \Delta(q^S_{H,H}) \geq 0.\end{aligned}$$

Together with $d\Gamma(k, q^O_{H,H})/dk > 0$ (see (A.6)), this inequality implies that S1's IC constraint is violated for all $k \in (\Delta(q^S_{H,H}), k_2)$. Overall, in this case, the PD does not exist for any value of $k$. □

# B Proofs of Section 4

The following three lemmas help to prove Proposition 3.

**Lemma B.5.** *(i) In the FD region,* $\frac{\partial \hat{c}_{FD}}{\partial q^O_0} < 0$ *and* $\frac{\partial \chi_{FD}}{\partial q^O_0} = 0$.

*(ii) In the PD region,* $\frac{\partial \hat{c}_{PD}}{\partial q^O_0} = 0$ *and* $\frac{\partial \chi_{PD}}{\partial q^O_0} < 0$.

*Proof of Lemma B.5. Proof of part (i).* By inspection of (6), recognizing the fact that $\partial q^O_H / \partial q^O_0 > 0$.

*Proof of part (ii).* From (17) one sees that $\hat{c}_{PD}$ is constant in $q^O_0$. As to $\chi_{PD}$, (A.3) and part (i) of this lemma imply

$$\frac{d\chi_{PD}}{dq^O_0} = \underbrace{\frac{\partial \chi_{PD}}{\partial \hat{c}_{FD}}}_{>0} \underbrace{\frac{\mathrm{d}\hat{c}_{FD}}{\mathrm{d}q^O_0}}_{<0} < 0.$$

□

**Lemma B.6.** *Suppose* $q^* \in (0, \bar{q}^O_0)$ *and define* $k^* \equiv k_2|_{q^O_0 = q^*} = k_3|_{q^O_0 = q^*}$. *We have:*

*(i)* $k_1 < k^*$;

*(ii)* $k_2 < k^* \iff q^O_0 > q^*$

*(iii)* $k_3 > k^* \iff q^O_0 > q^*$

*(iv)* $k_4 > k^* \iff q_0^O < q^*$

*Proof of Lemma B.6. Proof of part (i).* Inspection of (8) reveals that $\frac{\partial k_2}{\partial q_0^O} < 0$. Therefore, $k^* = k_2|_{q_0^O=q^*} > k_2|_{q_0^O=\bar{q}_0^O} = k_2|_{\hat{c}_{FD}=0} = \Delta(q_{H,H}^S) \geq k_1$.

*Proof of part (ii).* We have seen before that $\frac{\partial k_2}{\partial q_0^O} < 0$, hence $k^* = k_2|_{q_0^O=q^*} > k_2|_{q_0^O>q^*}$. The inequality is reversed for $q_0^O < q^*$.

*Proof of part (iii).* Inspection of (10) reveals that $\frac{\partial k_3}{\partial q_0^O} > 0$, hence $k^* = k_3|_{q_0^O=q^*} < k_3|_{q_0^O>q^*}$. The inequality is reversed for $q_0^O < q^*$.

*Proof of part (iv).* Recall from the proof of Proposition 2(a) that $\Gamma(k, q_0^O)$ crosses once the zero line from below at $k = k_4$. This fact, combined with

$$\frac{d\Gamma(k, q_0^O)}{dq_0^O} = -\underbrace{\frac{\partial k_3'}{\partial \chi_{PD}}}_{>0}\underbrace{\frac{d\chi_{PD}}{dq_0^O}}_{<0} > 0,$$

implies that $\frac{\partial k_4}{\partial q_0^O} < 0$. To complete the proof of the claim, observe that $k^* = k_4|_{q_0^O=q^*}$ because $\Gamma(k_2, q_0^O) = k_2 - k_3$ (see A.5) and $(k_2 - k_3)|_{q_0^O=q^*} = 0$. □

**Lemma B.7.** *Informativeness $\tau(H, H)$ satisfies the following properties.*

*(i) In the NIT region, $\tau(H, H) = 0$, which is constant in $q_{H,H}^O$ and $k$, and is minimal across all equilibrium configurations.*

*(ii) In the PD and FD regions, $\tau(H, H) = 1 - F(\hat{c}) + F(\hat{c})\chi$, with*

$$\frac{\partial \tau(H, H)}{\partial \chi} > 0 \text{ and } \left.\frac{\partial \tau(H, H)}{\partial \hat{c}}\right|_{\chi<1} < 0.$$

*(iii) In the ND region, $\tau(H, H) = 1$, which is constant in $q_{H,H}^O$ and $k$, and is maximal across all equilibrium configurations.*

*Proof of Lemma B.7. Proof of parts (i) and (iii).* Immediate.

*Proof of part (ii).* In the PD and FD equilibria, an S1 who observes the high signal reports truthfully with probability one. An S2 who also observes the high signal reports truthfully with probability one after no cancellation (which occurs with probability $1 - F(\hat{c})$) and with probability $\chi$ after cancellation (which occurs with probability $F(\hat{c})$). These considerations give the expression $\tau(H, H) = 1 - F(\hat{c}) + F(\hat{c})\chi$. The signs of the derivatives follow by inspection. □

We prove Proposition 3 first assuming that $q^* \in (0, \bar{q}_0^O)$ and then for the case $q^* = 0$.

*Proof of Proposition 3 when $q^* \in (0, \bar{q}_0^O)$. Proof of part (i).* Here, opposition is weak (strong) if $q_0^O > q^*$ ($q_0^O < q^*$).

**Case $k \leq k_1$.** Such values for $k$ belong to the ND region, and the boundary of this interval is independent of $q_0^O$. Further, in the ND region, speakers' and O's strategies are constant in $q_0^O$. Overall, informativeness is constant in $q_0^O$.

**Case $k \in (k_1, k^*)$, where $k^*$ is defined in Lemma B.6.** Such values for $k$ may belong to either the PD or the FD region. Note that $k_2|_{q_0^O = q^*} = k^* > k$ and $k_2|_{q_0^O = \bar{q}_0^O} = k_1 < k$. Because $k_2$ is decreasing in $q_0^O$, by continuity there exists a unique $q_k^* \in \left(q^*, \bar{q}_0^O\right)$ such that $k_2|_{q_0^O = q_k^*} = k$. By Lemma B.6, the following holds:

- $q_0^O < q^*$ implies that opposition is strong and $k \in (k_1, k_4)$, thus we have the PD equilibrium;
- $q_0^O \in [q^*, q_k^*)$ implies that opposition is weak and $k \in (k_1, k_2)$, thus we also have the PD equilibrium;
- $q_0^O \geq q_k^*$ implies that opposition is weak and $k \in [k_2, k_3)$, thus we have the FD equilibrium.

In each of these two regions, the comparative statics with respect to $q_0^O$ is as follows:

- In the PD region (i.e., $q_0^O \leq q_k^*$), increases in $q_0^O$ decrease informativeness: $\partial \hat{c}_{PD}/\partial q_0^O = 0$ and $d\chi_{PD}/dq_0^O < 0$ (see Lemma B.5(i));
- In the FD region (i.e., $q_0^O > q_k^*$), increases in $q_0^O$ increase informativeness: $\partial \hat{c}_{FD}/\partial q_0^O < 0$ and $\partial \chi_{FD}/\partial q_0^O = 0$ (see Lemma B.5(ii)).

**Case $k \geq k^*$.** Because $k_3$ is increasing in $q_0^O$, we have $k_3 \leq k_3|_{q_0^O = \bar{q}_0^O} \equiv \bar{k}_3$. Further, because $k_4$ is decreasing in $q_0^O$, we have $k_4 \leq k_4|_{q_0^O = 0} \equiv \bar{k}_4$.

If $k \geq \max\{\bar{k}_3, \bar{k}_4\}$, then $k$ belongs to the NIT region for all $q_0^O$, in which informativeness is constant in $q_0^O$.

Suppose next that $\bar{k}_3 < \bar{k}_4$ and $k \in (\bar{k}_3, \bar{k}_4)$. Because $k_4$ is decreasing in $q_0^O$ and $k_4|_{q_0^O = q^*} = k^*$, for each $k \in \left(k^*, \bar{k}_4\right)$ there exists a unique $q_{4,k}^* \in (0, q^*)$ such that $k_4|_{q_0^O = q_{4,k}^*} = k$. By Lemma B.6, the following holds:

- $q_0^O \leq q_{4,k}^*$ implies that opposition is strong and $k \in (k_1, k_4]$, thus we have the PD equilibrium;
- $q_0^O \in (q_{4,k}^*, q^*)$ implies that opposition is strong and $k > k_4$, thus we have the NIT equilibrium;
- $q_0^O \geq q^*$ implies that opposition is weak and $k > k_3$, thus we also have the NIT equilibrium.

We conclude that informativeness is weakly decreasing in $q_0^O$. The cutoff for $q_0^O$ can be taken to be $q_k^* = \bar{q}_0^O$.

Suppose now that $\bar{k}_4 < \bar{k}_3$ and $k \in (\bar{k}_4, \bar{k}_3)$. Because $k_3$ is increasing in $q_0^O$ and $k_3|_{q_0^O = q^*} = k^*$, for each $k \in \left(k^*, \bar{k}_3\right)$ there exists a unique $q_{3,k}^* \in (q^*, \bar{q}_0^O)$ such that $k_3|_{q=q_{3,k}^*} = k$. By Lemma B.6, the following holds:

- $q_0^O < q^*$ implies that opposition is strong and $k > k_4$, thus we have the NIT equilibrium;
- $q_0^O \in [q^*, q_{3,k}^*)$ implies that opposition is weak and $k > k_3$, thus we also have the NIT equilibrium;
- $q_0^O \geq q_{3,4}^*$ implies that opposition is weak and $k \in (k_2, k_3]$, thus we have the FD equilibrium.

We conclude that informativeness is weakly increasing in $q_0^O$. The cutoff for $q_0^O$ can be taken to be $q_k^* = 0$.

Last, suppose that $k \in \left(k^*, \min\{\bar{k}_3, \bar{k}_4\}\right]$. By Lemma B.6, the following holds:

- $q_0^0 \leq q_{4,k}^*$ implies that opposition is strong and $k \in (k_1, k_4]$, thus we have the PD equilibrium;
- $q_0^O \in (q_{4,k}^*, q^*)$ implies that opposition is strong and $k > k_4$, thus we have the NIT equilibrium;
- $q_0^O \in [q^*, q_{3,k}^*)$ implies that opposition is weak and $k > k_3$, thus we also have the NIT equilibrium;
- $q_0^O \geq q_{3,k}^*$ implies that opposition is weak and $k \in (k_2, k_3]$, thus we have the FD equilibrium.

We conclude that informativeness is decreasing in $q_0^O$ until $q_{4,k}^*$, stays constant in the interval $(q_{4,k}^*, q_{3,k}^*)$, and then is increasing from $q_{3,k}^*$. As for the point $q_k^*$ at which informativeness achieves its minimum, we can take any point in $(q_{4,k}^*, q_{3,k}^*)$.

*Proof of part (ii).*

**Case** $F(0) \in \left[\frac{\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)}{1+\Pr_S(\tilde{s}_2 = H|\tilde{s}_1 = H)}, \Pr_S\left(\tilde{s}_2 = H|\tilde{s}_1 = H\right)\right)$**.** Here, opposition is strong for all $q_0^O$.

All $k \leq k_1$ belong to the ND region for all $q_0^O$, in which equilibrium informativeness is constant in $q_0^O$ in the ND region.

Next, consider $k \in (k_1, \bar{k}_4)$. Observe that $\lim_{q_0^O \uparrow \bar{q}_0^O} k_4 = k_1$.[9] Because $k_4$ is decreasing in $q_0^O$, for each such $k$ there exists a $q_k^* \in (0, \bar{q}_0^O)$ such that $k = k_4|_{k=q_k^*}$. By Lemma B.6, the following holds:

- $q_0^0 \leq q_k^*$ implies that $k \in (k_1, k_4]$, thus we have the PD equilibrium;
- $q_0^0 > q_k^*$ implies that $k > k_4$, thus we have the NIT equilibrium.

[9] This claim follows from $\lim_{q_0^O \uparrow \bar{q}_0^O} k_2 = k_1 = \Delta(q_{H,H}^S)$ and $k_4 \in (k_1, k_2)$.

We conclude that informativeness is weakly decreasing in $q_0^O$.

Last, if $k \geq \bar{k}_4$, then $k$ belongs to the NIT region for all $q_0^O$, in which equilibrium informativeness is constant in $q_0^O$.

**Case** $F(0) \geq \Pr(\tilde{s}_2 = H | \tilde{s}_1 = H)$**.** Here, there are only the ND and NIT regions. In both of these regions, speakers' and O's strategies are constant in $q_0^O$. Also, the boundary between these regions is $k_1$, which is independent of $q_0^O$. We conclude that equilibrium informativeness is constant in $q_0^O$ for all $k$. □

*Proof of Proposition 3 when* $q^* = 0$. When $q^* = 0$, it is always the case that $k_2 < k_3$. Also, opposition is weak (strong) if $F(0) < \frac{\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}{1+\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}$ ($F(0) > \frac{\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}{1+\Pr_S(\tilde{s}_2=H|\tilde{s}_1=H)}$).

The proofs of part (ii) of the proposition (i.e., strong opposition) and part (i) for $k \leq k_1$ and $k \geq \bar{k}_3$ remain unchanged. Let us then consider the case of weak opposition with $k \in (k_1, \bar{k}_3)$. Define as $\bar{k}_2 \equiv k_2|_{q_0^O=0}$ the maximum of $k_2$ with respect to $q_0^O$ and as $\underline{k}_3 \equiv k_3|_{q_0^O=0}$ the minimum of $k_3$.

If $k \in (k_1, \bar{k}_2]$, as $q_0^O$ increases the equilibrium transitions from PD (when $k \in (k_1, k_2)$) to FD (when $k \in (k_2, k_3)$). Thus, informativeness is first decreasing and then increasing in $q_0^O$.

If $k \in (\bar{k}_2, \underline{k}_3)$, then we have the FD equilibrium for all $q_0^O$, in which case informativeness is increasing in $q_0^O$.

Last, if $k \in [\underline{k}_3, \bar{k}_3)$, as $q_0^O$ increases the equilibrium transitions from NIT (when $k > k_3$)) to FD (when $k \in (k_2, k_3)$). Thus, informativeness is weakly increasing in $q_0^O$. □

*Proof of Proposition 4.* Informativeness is maximal in the ND. Therefore, as $k$ increases and we transition out of the ND region, informativeness decreases.

Similarly, informativeness is minimal in the NIT. Hence, as $k$ increases and we transition into the NIT region, informativeness decreases.

In the FD region (if it exists), informativeness is constant in $k$, because $k$ affects neither speakers' disclosure strategies nor O's cancellation strategy.

There remains to show that informativeness is decreasing in $k$ in the PD region (if it exists). In the PD, informativeness is decreasing in $\hat{c}_{PD}$ and increasing in $\chi_{PD}$. Thus, to conclude the proof it suffices to observe that

$$\frac{\partial \hat{c}_{PD}}{\partial k} > 0 \text{ and } \quad \frac{d\chi_{PD}}{dk} = \underbrace{\frac{\partial \chi_{PD}}{\partial \hat{c}_{PD}}}_{<0} \underbrace{\frac{\partial \hat{c}_{PD}}{\partial k}}_{>0} < 0.$$

□

# C More general definition of informativeness

In this appendix, we provide an alternative notion of informativeness based on the mean-preserving spread ranking of DM's posterior.

Let $\tilde{\Theta} \equiv \mathrm{E}_{DM}(\tilde{\theta}|\tilde{r}_1, \tilde{d}_1, \tilde{r}_2, \tilde{d}_2)$ denote DM's posterior at the point in time when he has to decide which action to take. For a convex function $g(\cdot)$, define informativeness as the ex ante expectation

$$\mathcal{I} \equiv \mathrm{E}_{DM}\left(g(\tilde{\Theta})\right). \tag{C.1}$$

The expectation is calculated according to DM's prior. Because $g(\cdot)$ is convex, informativeness $\mathcal{I}$ increases if the distribution of the posterior $\tilde{\Theta}$ increases in a mean-preserving spread sense. This definition maps several notions of informativeness that have been used in decision theory, such as the mean squared error, posterior variance (Goldstein and Yang, 2017), and entropy (Sims, 2003). For example, when $g(\Theta) = \Theta^2$, maximization of $\mathcal{I}$ is equivalent to minimization of the mean squared error $\mathrm{E}\left(\left(\tilde{\theta} - \tilde{\Theta}\right)^2\right)$. And if $g(\Theta) = -\Theta(1-\Theta)$, informativeness $\mathcal{I}$ is equivalent to measuring the reduction in variance relative to the ex ante variance,

$$\mathrm{E}\left(\frac{\mathrm{Var}\left(\tilde{\theta}\right) - \mathrm{Var}\left(\tilde{\theta}|\tilde{r}_1, \tilde{d}_1, \tilde{r}_2, \tilde{d}_2\right)}{\mathrm{Var}\left(\tilde{\theta}\right)}\right) = 1 - \mathrm{E}\left(\frac{\tilde{\Theta}(1-\tilde{\Theta})}{q_0(1-q_0)}\right).$$

Last, minimizing the ex ante expectation of the entropy of the posterior distribution is the same as maximizing informativeness $\mathcal{I}$ for $g(\Theta) = \Theta\ln(\Theta) + (1-\Theta)\ln(1-\Theta)$.

To argue that our comparative statics results (Section 4) are qualitatively similar under the present definition of informativeness, we proceed as follows. First (Lemma C.1), we calculate informativeness in the different equilibrium regions. Second (Lemma C.2), we show that $\mathcal{I}$ as a function of $q^O_{H,H}$ and $k$ behaves in the same way as $\tau(H,H)$.

**Lemma C.1.** *Take informativeness $\mathcal{I}$ as defined in (C.1).*

- ***NIT.*** *In the NIT region, the public posterior at the end of the game is equal to the prior ($\tilde{\Theta} = q_0$), and thus informativeness equals*

$$\mathcal{I} = g(q_0) \tag{C.2}$$

*and it achieves its minimum.*

- ***PD and FD.*** *In the PD and FD, informativeness equals*

$$\begin{aligned}\mathcal{I} = &\Pr(\tilde{s}_1 = L)g(\Theta_L) \\ &+ (1 - F(\hat{c}))\left[\Pr(\tilde{s}_1 = H, \tilde{s}_2 = L)g(\Theta_{H,L}) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)g(\Theta_{H,H})\right] \\ &+ F(\hat{c})\left[\Pr(\tilde{s}_1 = H, \tilde{s}_2 = L) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)(1 - \chi)\right] g(\Theta'_{H,L}) \\ &+ F(\hat{c})\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)\chi g(\Theta_{H,H}),\end{aligned} \tag{C.3}$$

*where* $\Theta'_{H,L} \equiv \Pr\left(\tilde{\theta} = 1 | \tilde{r}_1 = H, \tilde{r}_2 = L\right)$*. The FD has* $\hat{c} > 0$ *and* $\chi = 0$*, whereas the PD has* $\hat{c} > 0$ *and* $\chi \in (0, 1)$*.*

- ***ND.*** *In the ND, informativeness equals*

$$\begin{aligned}\mathcal{I} = &\Pr(\tilde{s}_1 = L)g(\Theta_L) \\ &+ \Pr(\tilde{s}_1 = H, \tilde{s}_2 = L)g(\Theta_{H,L}) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)g(\Theta_{H,H}),\end{aligned} \tag{C.4}$$

*and it achieves its maximum.*

*Proof of Lemma C.1. Proof of part (i).* Immediate.

*Proof of part (ii).* Here, the public posterior at the end of the game is stochastic. With probability $\Pr(\tilde{s}_1 = L)$, S1 observes and truthfully reports $r_1 = L$, there is no subsequent information transmission, and the posterior is $\tilde{\Theta} = \Theta_L \equiv \mathrm{E}(\tilde{\theta}|\tilde{s}_1 = L)$. With the remaining probability, S1 observes and truthfully reports $r_1 = H$, and the terminal beliefs depend on whether O cancels S1. If O does not cancel S1, S2 is truthful and the posterior is $\tilde{\Theta} = \Theta_{L,s_2}$, where $\Theta_{s_1,s_2} \equiv \mathrm{E}(\tilde{\theta}|\tilde{s}_1 = s_1, \tilde{s}_2 = s_2)$ – this event occurs with probability $(1-F(\hat{c}))\Pr(\tilde{s}_1 = H, \tilde{s}_2 = s_2)$, where $\hat{c}$ is O's cancellation cutoff. If instead O cancels S1 and S2 observes $s_2 = H$, then S2 truthfully reports $r_2 = H$ with probability $\chi$, in which case the posterior is $\tilde{\Theta} = \Theta_{L,H}$ – this event occurs with probability $F(\hat{c})\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)\chi$. Last, if O cancels S1, and either S2 observes $s_2 = H$ and misreports, or S2 observes $s_2 = L$ and reports truthfully, then the posterior is

$$\begin{aligned}\Theta'_{H,L} &\equiv \Pr\left(\tilde{\theta} = 1 | \tilde{r}_1 = H, \tilde{r}_2 = L\right) \\ &= \frac{\Pr(\tilde{\theta} = 1, \tilde{s}_1 = H, \tilde{s}_2 = H)(1 - \chi) + \Pr(\tilde{\theta} = 1, \tilde{s}_1 = H, \tilde{s}_2 = L)}{\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)(1 - \chi) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = L)} \\ &= \frac{q_{H,H}(1 - \chi) + \frac{\Pr(\tilde{\theta}=1, \tilde{s}_1=H, \tilde{s}_2=L)}{\Pr(\tilde{s}_1=H, \tilde{s}_2=H)}}{(1 - \chi) + \frac{\Pr(\tilde{s}_1=H, \tilde{s}_2=L)}{\Pr(\tilde{s}_1=H, \tilde{s}_2=H)}},\end{aligned}$$

where to obtain the last expression we have divided both the numerator and denominator by $\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)$.

*Proof of part (iii).* In the ND, there is symmetric information if S1 reports $r_2 = H$ (because then S2 truthfully reports his signal regardless of cancellation), whereas communication stops after S1 reports $r_2 = L$. Informativeness is thus given by (C.3) evaluated at $\chi = 1$. □

**Lemma C.2.** *Take informativness $\mathcal{I}$ as defined in (C.1). For any convex function $g(\cdot)$, $\mathcal{I}$ satisfies the following properties.*

*(i) In the NIT region, informativeness is constant in $q^O_{H,H}$ and $k$, and it is minimal across all equilibrium configurations.*

*(ii) In the PD and FD regions,*

$$\frac{\partial \mathcal{I}}{\partial \chi} < 0 \text{ and } \left.\frac{\partial \mathcal{I}}{\partial \hat{c}}\right|_{\chi<1} > 0.$$

*(iii) In the ND region, informativeness is constant in $q^O_{H,H}$ and $k$, and it is maximal across all equilibrium configurations.*

*Proof of Lemma C.2. Proof of part (i).* Inspection of (C.2) shows that informativeness is constant in the NIT region. The fact that here informativeness is minimal follows from Jensen's inequality: the posterior is degenerate and $g(\cdot)$ is convex.

*Proof of part (ii).* In the PD and FD regions, the derivative of informativeness with respect to $\chi$ is

$$\begin{aligned}\frac{\partial \mathcal{I}}{\partial \chi} &= F(\hat{c})\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)\left[-g(\Theta'_{H,L}) + g(\Theta_{H,H})\right] \\ &\quad + F(\hat{c})\left[\Pr(\tilde{s}_1 = H, \tilde{s}_2 = L) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)(1-\chi)\right] g'(\Theta'_{H,L})\frac{\partial \Theta'_{H,L}}{\partial \chi} \\ &= F(\hat{c})\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)\left[-g(\Theta'_{H,L}) + g(\Theta_{H,H}) - g'(\Theta'_{H,L})(\Theta_{H,H} - \Theta'_{H,L})\right] > 0,\end{aligned} \tag{C.5}$$

where for the last expression we have used

$$\frac{\partial \Theta'_{H,L}}{\partial \chi} = -(\Theta_{H,H} - \Theta'_{H,L})\frac{\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)}{\Pr(\tilde{s}_1 = H, \tilde{s}_2 = L) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)(1-\chi)}.$$

The signs of the derivative in (C.5) follows from convexity of $g(\cdot)$. Next, the derivative of informativeness with respect to $\hat{c}$ is

$$\begin{aligned}\left.\frac{\partial \mathcal{I}}{\partial \hat{c}}\right|_{\chi<1} &= -f(\hat{c})\left[\Pr(\tilde{s}_1 = H, \tilde{s}_2 = L)g(\Theta_{H,L}) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)g(\Theta_{H,H})\right] \\ &\quad + f(\hat{c})\left[\Pr(\tilde{s}_1 = H, \tilde{s}_2 = L) + \Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)(1-\chi)\right] g(\Theta'_{H,L}) \\ &\quad + f(\hat{c})\Pr(\tilde{s}_1 = H, \tilde{s}_2 = H)\chi g(\Theta_{H,H}) < \left.\frac{\partial \mathcal{I}(\hat{c},\chi)}{\partial \hat{c}}\right|_{\chi=1} = 0,\end{aligned} \tag{C.6}$$

where the inequality follows from (C.5).

*Proof of part (iii).* Inspection of (C.4) shows that informativeness is constant in the ND region. To show that informativeness is maximal here, recall that, in the ND, informativeness is given by (C.3) evaluated at $\chi = 1$. Then, part (ii) of this lemma shows that $\mathcal{I}$ is maximized at $\chi = 1$. □